

# LAPORAN DATA PRE-PROCESSING DENGAN PYTHON

ARIF FRIMA ARI SUWADJI - 221011700443 - 03SIFE003

## DATA PRE PROCESSING - FILTERING

INPUT	OUTPUT												
<code>pwd</code>	<pre>In [5]: pwd  Out[5]: 'C:\\Users\\arifs.DESKTOP-EQJN52\\Documents\\unpam\\Semester3\\Pengantar_Big_Data\\KelasBigData'</pre>												
<code>import pandas as pd from string import punctuation import nltk nltk.download('stopwords')</code>	<pre>In [56]: import pandas as pd from string import punctuation import nltk nltk.download('stopwords')  [nltk_data] Downloading package stopwords to C:\\Users\\arifs.DESKTOP- [nltk_data] EQJN52\\AppData\\Roaming\\nltk_data... [nltk_data] Unzipping corpora\\stopwords.zip.  Out[56]: True</pre>												
<pre>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</pre>	<pre>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]:</pre> <table><thead><tr><th></th><th>data</th></tr></thead><tbody><tr><td>0</td><td>This is introduction to NLP</td></tr><tr><td>1</td><td>It is likely to be useful, to people</td></tr><tr><td>2</td><td>Machine learning is the new electricity</td></tr><tr><td>3</td><td>There would be less hype around AI nd more act...</td></tr><tr><td>4</td><td>python is the best tool!</td></tr></tbody></table> <pre>In [50]: df.index/(len(df)-1)</pre>		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!
	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
<pre>stopwords = set(nltk.corpus.stopwords.words('english')) def proses(kalimat):     kalimat = kalimat.lower()     kalimat = nltk.tokenize.word_tokenize(kalimat)     kalimat = [kata for kata in kalimat if kata not in stopwords]     return kalimat  hasil = df['data'] hasil = hasil.apply(proses) hasil</pre>	<pre>In [59]: stopwords = set(nltk.corpus.stopwords.words('english')) def proses(kalimat):     kalimat = kalimat.lower()     kalimat = nltk.tokenize.word_tokenize(kalimat)     kalimat = [kata for kata in kalimat if kata not in stopwords]     return kalimat  In [60]: hasil = df['data'] hasil = hasil.apply(proses) hasil  Out[60]: 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</pre>