Analytics Vidhya

Tokenization

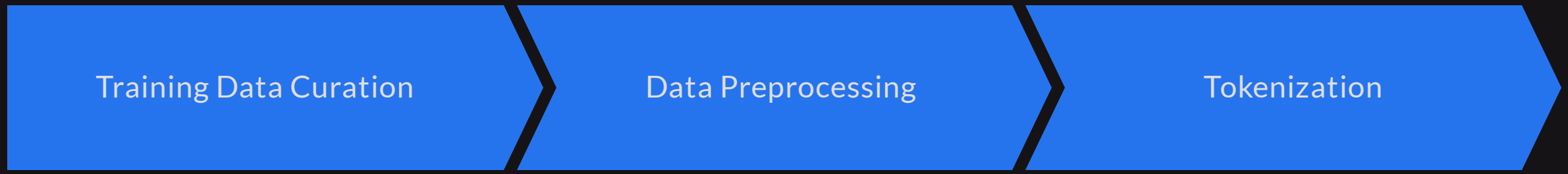
Instructor

Sourab Mangrulkar

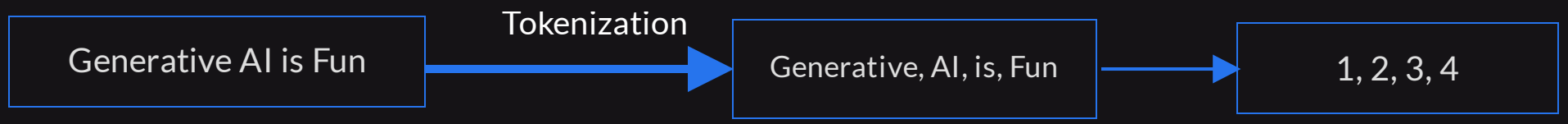
Machine Learning Engineer at

Creator of PEFT

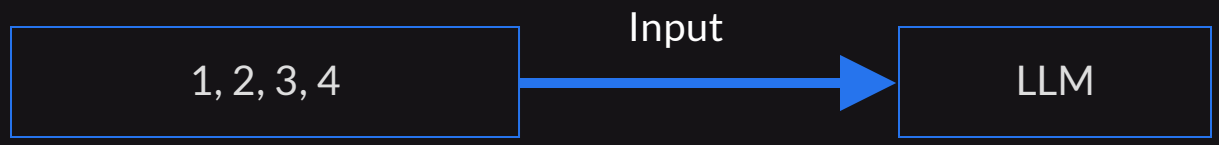
Recap



What is Tokenization?



Why Tokenization?



Steps involved in tokenization

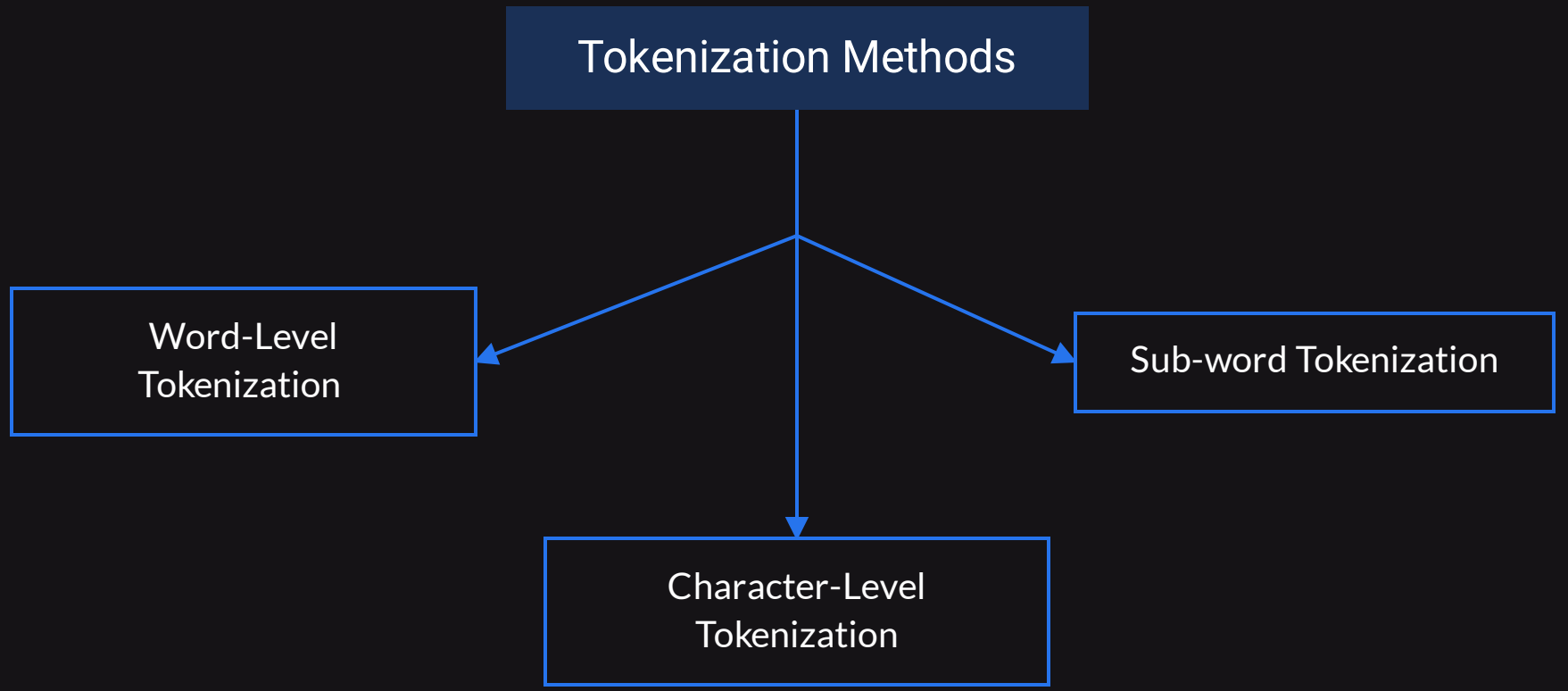
- Train tokenizer

Steps involved in tokenization

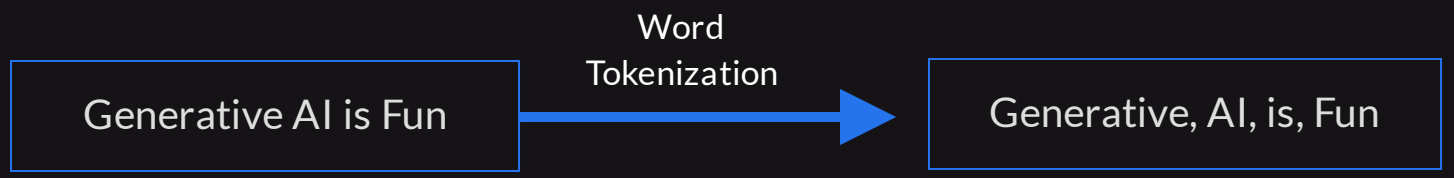
- Train tokenizer
- Apply tokeniser on training data

Types of Tokenization

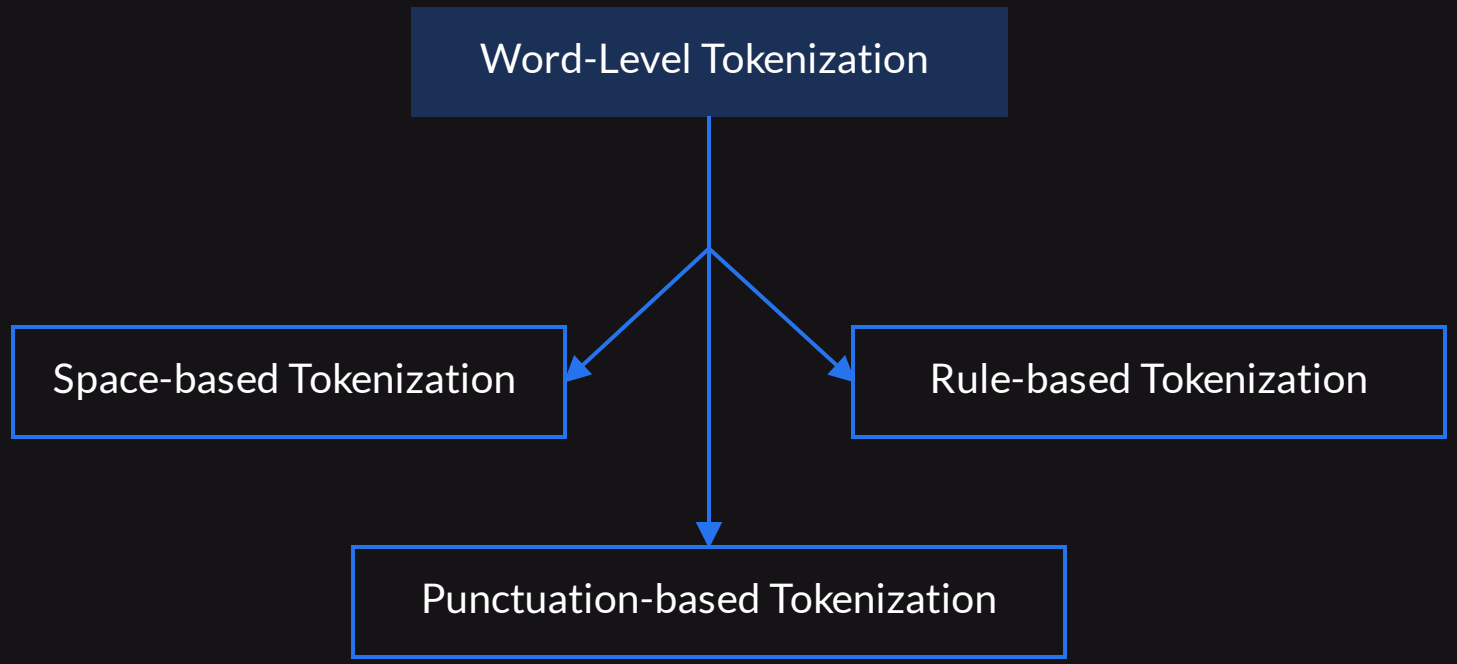
Tokenization Methods



Word Level Tokenization?



Word Level Tokenization



Pros and Cons

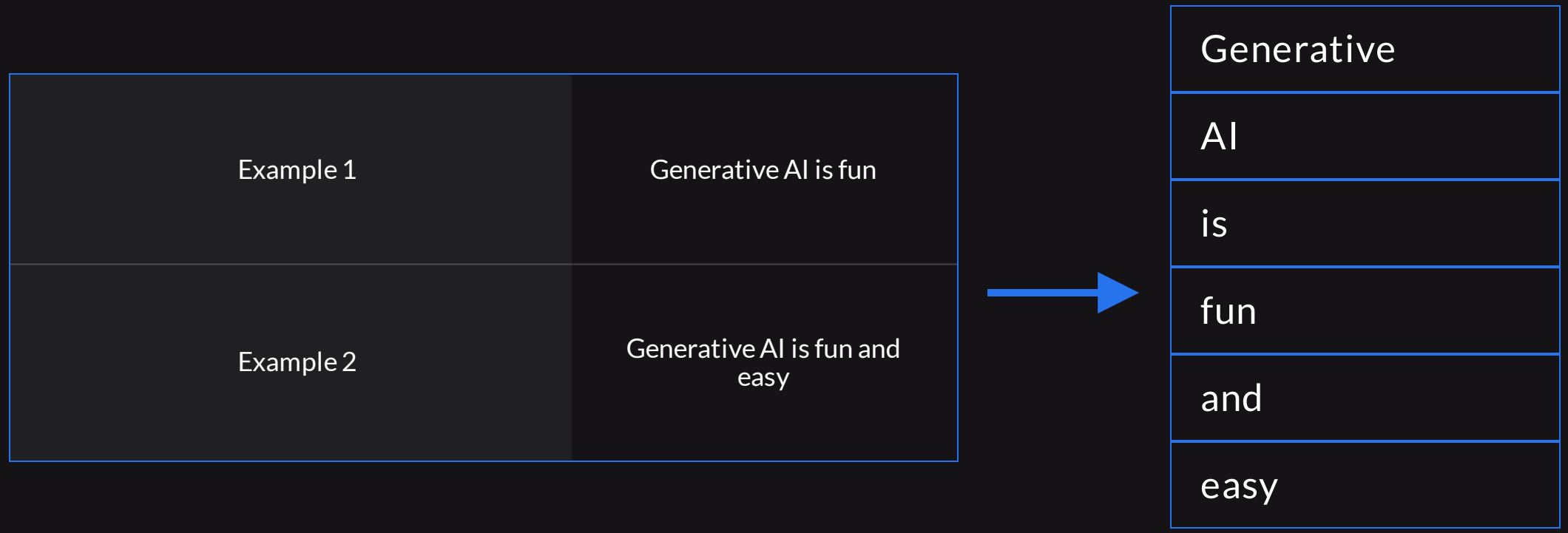
Pros

- Simple and Efficient

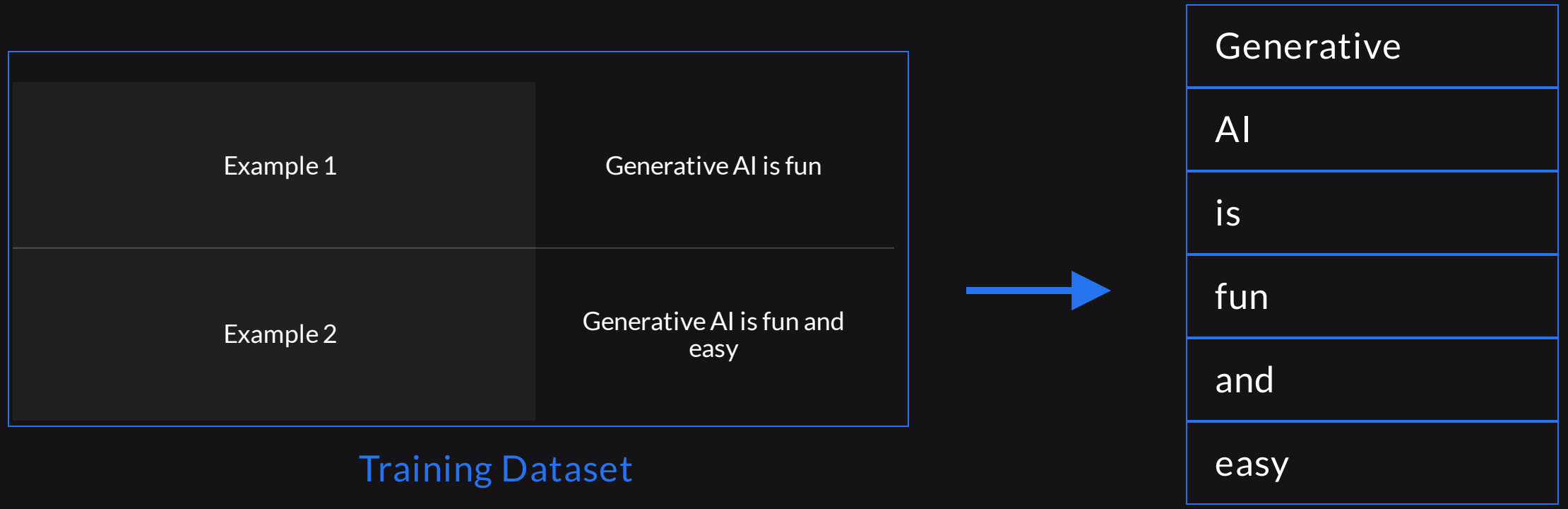
Cons

- Massive Text Vocabulary
- Out of Vocabulary

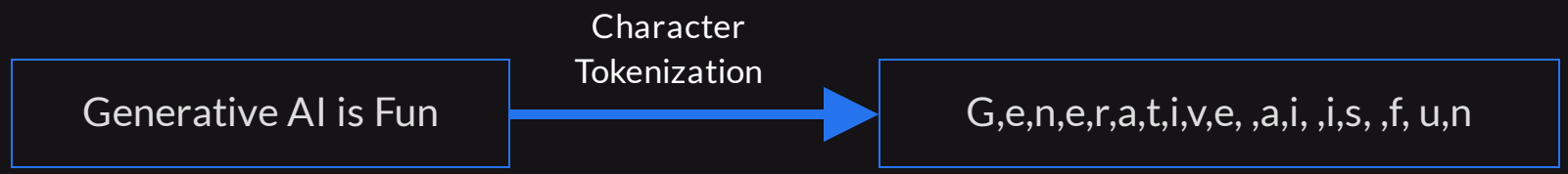
OOV



OOV



Character-Level Tokenization



Pros and Cons

Pros

Limited Vocabulary

Handle OOV

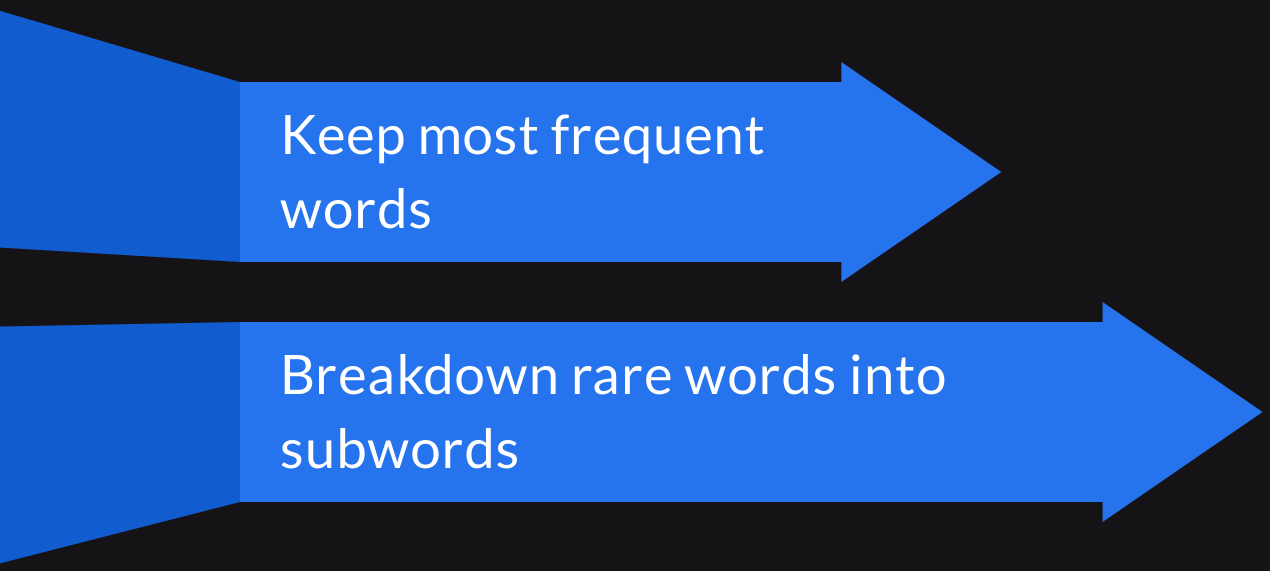
Cons

Learning meaningful input representations

Generates longer sequences

Sub Word Tokenization

Sub Word Tokenization



Different Methods: Sub Word Tokenization

Byte Pair Encoding

1 GPT-2

2 GPT-3

Byte Pair Encoding

Creates tokens by merging frequently occurring characters

Working of Byte Pair Encoding

Sample Corpus

low : 5

lower : 2

newest : 6

widest :3

Step 1: Add special character </w>

Sample Corpus

low</w> : 5

lower</w> : 2

newest</w> : 6

widest</w> :3

Step 2: Tokenise words into individual characters

Sample Corpus

l o w </w> : 5

l o w e r </w> : 2

n e w e s t </w> : 6

w i d e s t </w> :3

Step 3: Initialize the vocabulary

Sample Corpus

l o w </w> : 5

l o w e r </w>: 2

n e w e s t </w> : 6

w i d e s t </w>:3

Token
</w>
l
o
w
e
r
n
s
t
i
d

Analytics Vidhya

Step 4: Find the most frequent pair

Sample Corpus

l o w </w> : 5

l o w e r </w>: 2

n e w e s t </w>: 6

w i d e s t </w> :3

Token
</w>
l
o
w
e
r
n
s
t
i
d

Step 5: Merge the pair

Sample Corpus

l o w </w>: 5

l o w e r</w> : 2

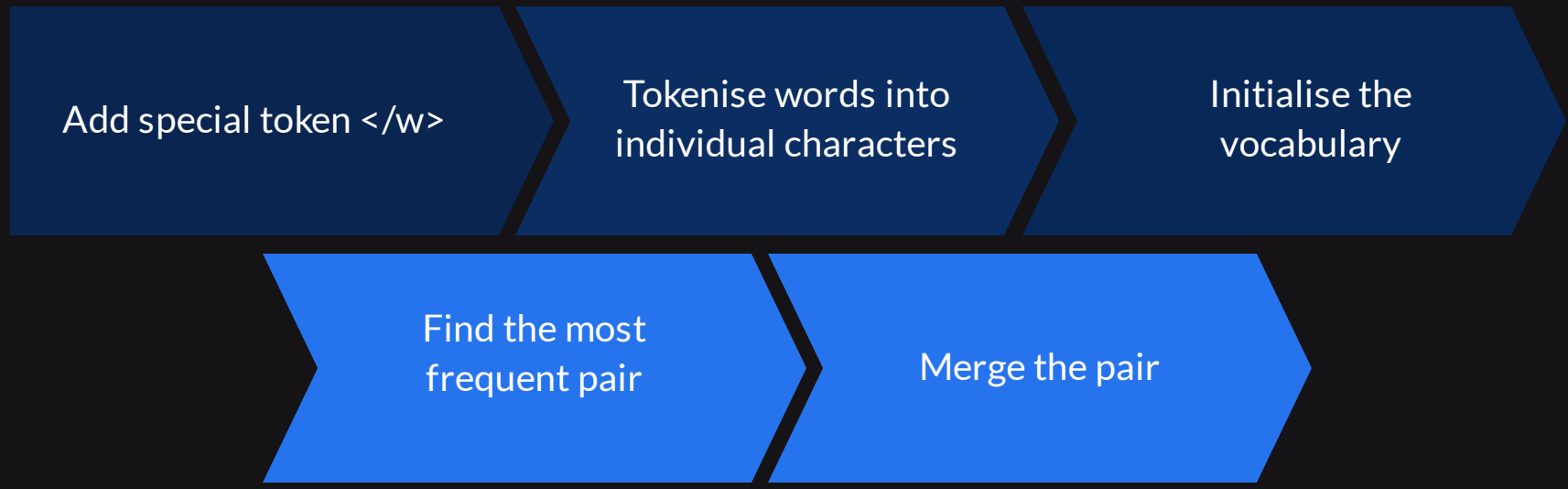
n e w es t</w> : 6

w i d es t</w> :3

Token
</w>
l
o
w
e
r
n
s
t
i
d
es

Analytics Vidhya

Working of Byte Pair Encoding



Iteration 2: Find the most frequent pair

Sample Corpus

l o w </w> : 5

l o w e r</w>: 2

n e w es t</w>: 6

w i d es t</w> :3

Token
</w>
l
o
w
e
r
n
s
t
i
d
es

Iteration 2: Merge the pair

Sample Corpus

l o w </w> : 5

l o w e r</w> : 2

n e w est</w> : 6

w i d est</w>:3

Token
</w>
l
o
w
e
r
n
s
t
i
d
es
est

After 5 Iterations

Sample Corpus

low </w> : 5

low e r</w>: 2

n e w est</w>: 6

w i d est</w> :3

Token
</w>
l
o
w
e
r
n
s
t
i
d
es
est
est</w>
lo
low

Steps involved in tokenization

- Train tokenizer
- Apply tokeniser on training data

Analytics Vidhya

Applying BPE Tokenizer

lowest -> low,est

Token
</w>
l
o
w
e
r
n
s
t
i
d
es
est
est</w>
lo
low

Thank You

Analytics Vidhya