

N-gram이란 간단하게 말해서 입력한 문자열을 N개의 기준 단위로 절단하는 방법이다. 예를 들어 "Here is a dog" 라는 문장을 문자 절단 단위의 3-gram으로 만든다면,

"Her", "ere", "re_", "e_i", "_is", "is_", "s_a", "_a_", "a_d", "_do", "dog" 로 만들 수 있다.

같은 작업을 단어 단위의 2-gram으로 만든다면,

"Here is", "is a", "a dog" 가 된다.

This is Big Data AI Book

Uni-Gram	This	Is	Big	Data	AI	Book
Bi-Gram	This is	Is Big	Big Data	Data AI	AI Book	
Tri-Gram	This is Big	Is Big Data	Big Data AI	Data AI Book		

Naive Next word prediction

- how are you doing
- how are you
- how are they

Trigram	Count
how are you	2
are you doing	1
how are they	1

Naive Next word prediction

how are you

Trigram	Count
how are you	2
are you doing	1
how are they	1

Naive Spell checker

- quality
- quater
- quit

Bigram	Count
qu	3
ua	2
al	1
li	1
it	2
ty	1
at	1
te	1
er	1
ui	1

Naive Spell checker

qwal

qual

Bigram	Count
qu	3
ua	2
al	1
li	1
it	2
ty	1
at	1
te	1
er	1
ui	1