

N-gram 언어 모델

- 통계기반 언어모델의 일종. SLM과 같이 카운트 기반 통계적 접근을 사용한다.
- 전통적 SLM과 달리 이전에 등장한 모든 단어가 아닌 **일부 단어만 고려**하는 방법을 사용한다.
 - n-gram에서의 n은 코퍼스 내 단어들을 n개씩 묶어서 빈도를 학습했음을 의미한다.
- **이전 n-1개의 단어를 보고 n번째 단어를 예측**하는 방식
- 임의의 개수만큼의 이전 단어만 참고하여 확률을 근사
 - 코퍼스에서 해당 단어시퀀스를 카운트할 확률이 높아진다.

N-gram 언어 모델

표현	빈도
영원히	104
기억될	29
최고의	3503
명작이다	298
영원히 기억될	7
기억될 최고의	1
최고의 명작이다	23
기억될 최고의 명작이다	17
영원히 기억될 최고의 명작이다	0

- 영원히 기억될 최고의 시퀀스 뒤에 '명작이다' 라는 단어가 올 확률을 **trigram**으로 근사해보면 얼마일까?