

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?

N-gram 언어모델의 한계

- 희소문제(Sparsity problem)
 - 카운트 기반 접근방식의 본질적 한계
 - 코퍼스 내에 단어시퀀스가 없을 확률은 여전히 존재
- n의 선택은 trade-off
 - n크기를 키우면 : 예측 정확도 상승 but 희소문제 증가, 모델사이즈 증가
 - n크기를 줄이면 : 희소문제 감소 but 예측 정확도 감소
- long-term dependency : 정해진 개수의 이전 토큰만을 반영하므로 고려할 수 있는 시퀀스 범위 한정됨. >> 모델의 정확도와 연관