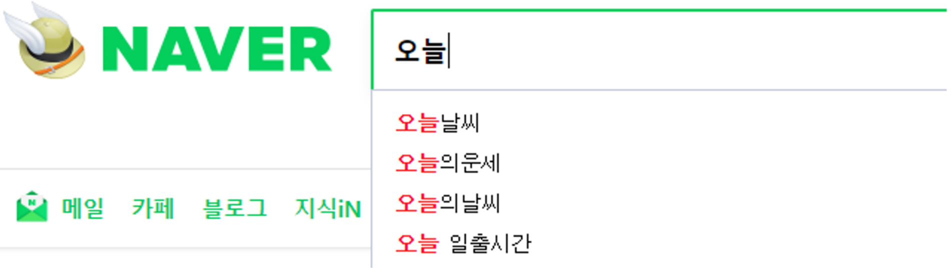
N-gram
Language Model

Language Modeling

Language Model은 다음 단어가 무엇일지를 예측하는 일!

- 이전의 단어 w_1, w_2, w_3, …, w_t가 주어졌을 때, 다음 단어 w_{t+1}의 확률은?

the students opened their _______



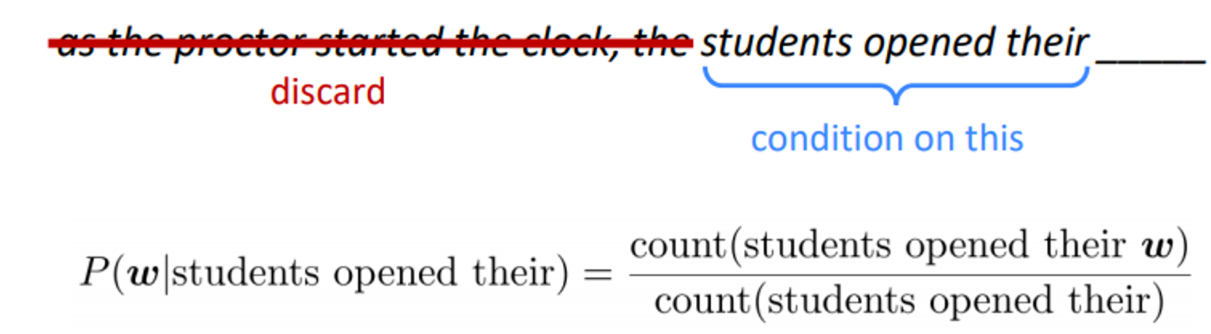
오늘

오늘날씨
오늘의운세
오늘의날씨
오늘 일출시간

N-gram

n-gram 언어 모델의 가정: 다음 단어의 확률은 이전 *n-1*개 단어에 의존한다.

4그램 언어 모델의 예시:



~~*as the proctor started the clock, the*~~ *students opened their* ______

discard

condition on this

$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\text{count}(\text{students opened their } \boldsymbol{w})}{\text{count}(\text{students opened their})}$$

- “students opened their” 1000번 등장
- “students opened their books” 400번 등장
  - $P(books|students\ opened\ their) = 0.4$
- “students opened their exams” 100번 등장
  - $P(exams|students\ opened\ their) = 0.1$

N-gram

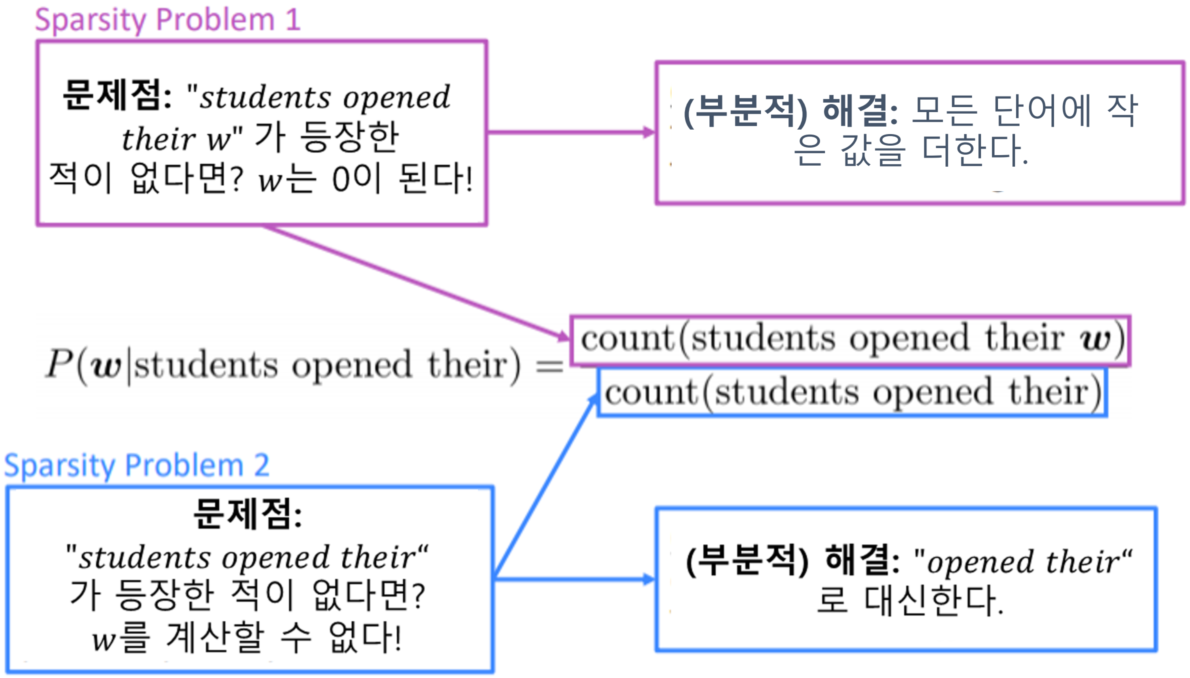
“You are uniformly charming!” cried he, with a smile of associating and now and then I bowed and they perceived a chaise and four to wish for.

제인 오스틴 3-gram 모델로 만든 랜덤한 문장

I 'll admit there 's the way many americans would find severe. Second the world 's major stock markets continued their rise the agriculture department was set just before the house in N and a number of studies suggest that while technology offers almost endless hope when to do something about it now looks like a duck.

PTB 데이터셋에서 2-gram 모델로 만든 문장

N-gram



Sparsity Problem 1

문제점: "*students opened their w*" 가 등장한 적이 없다면? *w*는 0이 된다!

(부분적) 해결: 모든 단어에 작은 값을 더한다.

$$P(\boldsymbol{w}|\text{students opened their}) = \frac{\boxed{\text{count(students opened their } \boldsymbol{w})}}{\boxed{\text{count(students opened their)}}}$$

Sparsity Problem 2

문제점: "*students opened their*“ 가 등장한 적이 없다면? *w*를 계산할 수 없다!

(부분적) 해결: "*opened their*“ 로 대신한다.

<u>Note:</u> Increasing *n* makes sparsity problems *worse*. Typically we can’t have *n* bigger than 5.

다른 문제점: Tom was watching TV in his room. Mary came into the room. Mary said hi to □ ?