N-gram과 라플라스 스무딩

1. N-gram의 한계: Zero 확률 문제

N-gram 모델은 주어진 (n-1)개의 단어가 등장한 후, 특정 단어가 등장할 확률을 통계적으로 계산합니다. 하지만 훈련 데이터(코퍼스)에 등장하지 않은 단어 조합이 실제 문장에서는 나타날 수 있습니다.
이 경우, 해당 N-gram의 확률은 0이 되어 전체 문장 확률이 0이 되는 문제가 발생합니다.
이러한 현상을 **Zero 확률 문제** 또는 **희소성 문제(Sparse Data Problem)** 라고 합니다.

예시)

- 코퍼스에 "나는 밥을 먹는다"는 있지만, "나는 밥을 마신다"는 없다면,
  P(마신다 | 나는 밥을) = 0

2. 라플라스 스무딩(Laplace Smoothing, Add-one Smoothing)

Zero 확률 문제를 해결하기 위해 **라플라스 스무딩**을 사용합니다.
라플라스 스무딩은 모든 가능한 N-gram 조합에 1을 더해줌으로써, 등장하지 않은 조합에도 최소한의 확률을 부여합니다.

수식

- 기존 확률:
  $P(w_n \mid w_1, ..., w_{n-1}) = Count(w_1, ..., w_n) / Count(w_1, ..., w_{n-1})$
- 라플라스 스무딩 적용:
  $P(w_n \mid w_1, ..., w_{n-1}) = (Count(w_1, ..., w_n) + 1) / (Count(w_1, ..., w_{n-1}) + V)$
  여기서 V는 전체 단어(어휘)의 개수입니다.

예시

코퍼스: "나는 밥을 먹는다", "너는 밥을 먹는다"

- "밥을 먹는다"의 등장 횟수: 2
- "밥을 마신다"의 등장 횟수: 0
- 어휘(V): 5 (나는, 너는, 밥을, 먹는다, 마신다)

라플라스 스무딩 적용 후

- P(먹는다 | 밥을) = (2 + 1) / (2 + 5) = 3/7
- P(마신다 | 밥을) = (0 + 1) / (2 + 5) = 1/7

즉, 등장하지 않은 조합에도 0이 아닌 작은 확률이 부여되어, 실제 문장 생성이나 확률 계산에서 0 확률로 인해 전체 결과가 무의미해지는 것을 방지할 수 있습니다.

3. 라플라스 스무딩의 한계

- 모든 조합에 동일하게 1을 더하기 때문에, 어휘 수(V)가 클수록 실제로 등장한 조합의 확률이 과도하게 낮아질 수 있습니다.
- 실제로는 등장하지 않을 법한 조합에도 확률이 부여됩니다.

이러한 한계를 보완하기 위해 Add-k 스무딩(k를 1보다 작은 값으로 설정)이나 Good-Turing 추정 등 다양한 스무딩 기법이 존재합니다.