1. 통계적언어모델 이란

통계적 언어 모델(Statistical Language Model)은 자연어 문장이나 단어의 시퀀스에 확률을 할당하는 모델입니다. 이는 특정 맥락에서 단어나 문장이 나타날 확률을 계산하여, 기계번역, 음성 인식, 맞춤법 검사, 문장 생성 등 다양한 자연어처리 작업에 활용됩니다. 단어의 빈도와 순서에 대한 통계적 정보를 학습하여 언어의 패턴을 모델링합니다.

2. N-gram 이란

N-gram은 연속된 N개의 항목(단어, 문자, 음절 등)을 하나의 단위로 처리하는 모델입니다.

- Unigram(1-gram): 단일 단어/문자만 고려 (예: "나", "는", "밥")
- Bigram(2-gram): 연속된 2개 단어/문자 (예: "나는", "는 밥", "밥을")
- Trigram(3-gram): 연속된 3개 단어/문자 (예: "나는 밥", "는 밥을", "밥을 먹는다")

N-gram 모델은 이전 n-1개 단어만을 고려하여 다음 단어의 확률을 예측하는 마르코프 가정(Markov assumption)을 기반으로 합니다. 간단하고 구현이 쉬우면서도 효과적이라는 장점이 있지만, 희소성 문제(sparse data problem)와 문맥 범위 제한이라는 단점도 있습니다.

3. Log처리 하는 이유?

자연어 처리에서 확률 계산 시 로그 함수를 사용하는 주요 이유는 다음과 같습니다:

1. 수치적 안정성(Numerical Stability): 확률값은 매우 작은 수(0~1 사이)이므로, 여러 확률을 곱하면 언더플로우(underflow) 문제가 발생할 수 있습니다. 로그를 취하면 곱셈이 덧셈으로 변환되어 이 문제를 해결할 수 있습니다.
   - $P(A)\times P(B)\times P(C) \rightarrow \log(P(A)) + \log(P(B)) + \log(P(C))$
2. 계산 효율성: 로그를 사용하면 곱셈 연산이 덧셈으로 바뀌어 계산 효율이 향상됩니다.
3. 확률 값의 비교: 매우 작은 확률값들을 비교할 때 로그 스케일에서는 차이가 더 명확하게 드러납니다.
4. 학습 안정성: 딥러닝에서 손실 함수로 자주 사용되는 cross-entropy는 로그 확률을 기반으로 하며, 경사 하강법에서 안정적인 학습을 가능하게 합니다.

로그 확률은 항상 음수값을 가지며(0~1 사이의 확률의 로그이므로), 값이 0에 가까울수록(즉, 로그값이 0에 가까울수록) 더 높은 확률을 의미합니다.