[숙제11] N-그램 언어 모형

언어와 컴퓨터

2018년 11월 23일 금요일 13시까지

- 보고서 파일 `hw11_00000.pdf`를 제출하라.
- 보고서에는 계산 과정·결과 및 간단한 설명이 있으면 된다. 계산 과정에서는 파이썬 대화형 모드 인터프리터를 계산기처럼 사용하면 된다.

1 N-그램 언어 모형

1.0 데이터

훈련 코퍼스의 단어 수는 $N = 1410000000$, 유니그램·바이그램·트라이그램의 빈도는 아래와 같다고 하자.

유니그램	빈도	바이그램	빈도	트라이그램	빈도
"하늘은"	3520000	"하늘은 파랗고"	56100	"하늘은 파랗고 단풍잎은"	34
"파랗고"	392000	"파랗고 단풍잎은"	23	"파랗고 단풍잎은 빨갛고"	0
"단풍잎은"	34600	"단풍잎은 빨갛고"	160	"단풍잎은 빨갛고 은행잎은"	3
"빨갛고"	339000	"빨갛고 은행잎은"	85	"빨갛고 은행잎은 노랗고"	85
"은행잎은"	24300	"은행잎은 노랗고"	198		
"노랗고"	359000				

실험 집합은 $W = w_1w_2w_3w_4w_5w_6 =$ "하늘은 파랗고 단풍잎은 빨갛고 은행잎은 노랗고"이다.

1.1 모형별 확률 계산 (3점)

확률 $P(W)$을 유니그램·바이그램·트라이그램 모형으로 각각 추정하라.

주의사항

1. N-그램 모형에서는 문장 앞에 문장 시작 표시 `<s>`를 $(N-1)$ 개 넣는다.
2. 이 문제에서는 문장 시작 바이그램과 트라이그램의 확률이 아래와 같다고 가정하라.
   - P("하늘은"|`<s>`) $=$ P("하늘은"|`<s><s>`) $=$ P("하늘은")
   - P("파랗고"|`<s>` "하늘은") $=$ P("파랗고"|"하늘은")

1.2 모형별 복잡도 평가 (3점)

유니그램·바이그램·트라이그램 모형에서 복잡도 $PP(W)$를 각각 계산하고 그 결과를 비교하라.

1.3 보간법 (4점)

트라이그램 모형에서 확률의 추정치 $\hat{P}(w_n|w_{n-2}w_{n-1})$를 아래와 같이 새로 정의하자.

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1 P(w_n|w_{n-2}w_{n-1}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n), \quad w = 1, 2, 3, \cdots, 6$$

이렇게 정의된 $\hat{P}(w_n|w_{n-2}w_{n-1})$를 사용한 새로운 모형에서 $W = w_1w_2w_3\cdots w_6$의 확률과 복잡도를 구하려고 한다. $(\lambda_1, \lambda_2, \lambda_3) = (0.5, 0.3, 0.2)$ 일 때와 $(\lambda_1, \lambda_2, \lambda_3) = (0.7, 0.2, 0.1)$일 때의 $P(W)$와 $PP(W)$를 각각 계산하라.

1.4 기타 (+2점)

계산 과정에서 함수를 정의하거나 NumPy를 사용하면 가산점을 받을 수 있다.