2강 연습문제

1. 다음 말뭉치(corpus)에 대해서 윈도우 크기를 1과 2로 잡았을 때 교재 코드를 사용하여 PPMI행렬을 각각 출력하시오.

   `The sky is very blue and the sky is very beautiful today.`

2. 확률 변수 X, Y의 확률분포가 다음과 같이 주어져 있다.

Y \ X	1	2	3	4
1	1/8	1/16	1/32	1/32
2	1/16	1/8	1/32	1/32
3	1/16	1/16	1/16	1/16
4	1/4	0	0	0

   (i) PMI 테이블과 PPMI 테이블을 구하시오.

   (ii) 교재의 PPMI 코드는 대칭행렬을 가정하고 작성된 코드이다. 일반적인 행렬에도 적용되도록 코드를 수정한뒤 (i)의 결과를 검산하시오.

3. `count_method_big.py`를 기반으로 다음 코드를 작성하시오.

   (i) 동시발생행렬과 PPMI행렬 계산결과를 pickle파일에 저장하시오.

   (ii) (i)의 pickle파일을 불러온 후 PPMI행렬 대신 동시발생행렬에 대하여 SVD를 적용한 후 `you, year, car, toyota`와 코사인 유사도가 가장 높은 5개의 단어를 출력하시오. 기존 결과와 비교하고 이유를 해석하시오.

   (iii) (i)의 pickle파일을 불러온 후 벡터표현의 차원을 각각 5, 20, 100으로 잡았을 때 `you, year, car, toyota`와 코사인 유사도가 가장 높은 5개의 단어를 각각 출력하시오. 기존 결과와 비교하고 이유를 해석하시오.