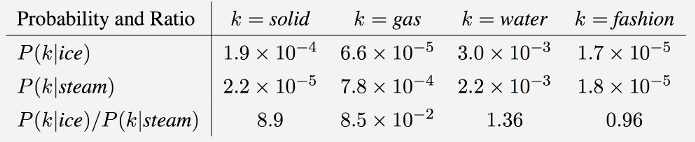
**GloVe**

Word2vec에서는 내적을 해서 분자의 likelihood를 높였는데 이는 코사인 유사도와 관련이 있다. = 임베딩된 단어간 유사도도 내적을 통해 한다.  
GloVe에서는 단어-문맥 행렬에 특이값분해(SVD) 실시 => 데이터 차원 축소  
노이즈를 줄여 내재적 의미 이끄는 LSA의 단점 또한 언급 => GloVe가 LSA, Word2Vec 비판

LSA는 말뭉치 전체의 통계적인 정보 모두 활용 => but 단어/문서 간 유사도 측정 어려움  
W2V은 지정한 윈도우에서만 학습/분석 => 말뭉치 전체의 co-occurrence(공기정보) 정보 반영 어려움

임베딩된 두 단어벡터의 내적이 말뭉치 전체에서 동시 등장 확률 로그값이 되도록 objective function 설정 = “단어벡터간 유사도 측정 수월, 말뭉치 전체의 통계 정보를 좀 더 잘 반영해보자”

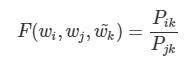
동시에 같이 등장한 단어의 빈도를 각각 세어서 말뭉치의 단어 개수로 나눠줌 = 동시등장확률 = the words’ probability of co-occurrence)



Steam 등장 시 solid 나올 확률 > ice 등장 시 solid 나올 확률 (비율= 8.9)

특정 단어 k가 주어졌을 때 두 단어(steam, ice) 처럼의 내적 값 = 두 단어의 동시등장확률 간 비율 이 되게끔 임베딩이 목적 => solid가 주어졌을 때 ice와 steam 벡터 내적 => 8.9

K 자리에 ice, steam이 올 수 있고, 전제 자리에 solid가 올 수도 있다. => 이거를 전체 단어에 대해 = 논문 아이디어

이 식을 만족하는 임의의 함수 F를 찾고자 함.

P-ik = P(k|i) = i번째 단어 주변에 k번째 단어가 등장할 조건부 확률 = i번째 단어와 k번째 단어가 동시에 등장한 빈도수 (X-ik)를 로 나눈 값 <위에서 여태 이야기한 것>