과제08 EM Clustering

첨부된 데이터 파일 `emcluster_sample.txt`를 이용하여 clabel 정보는 없다고 가정하고 clustering을 하여라. 집단은 2개라고 가정하여라. Naive Bayes Model을 이용하여라.

(데이터) 다음과 같이 문서 고유 번호(docid), 해당 문서에 들어 있는 단어의 목록 (text), 해당 문서의 부류(clabel)로 이루어진 자료가 있다.

```
docid   text    clabel
1       1 2 4   2
2       2 3     2
3       3 4     1
4       4       1
```

이 자료를 다음과 같은 형식으로 정리할 수 있다.

```
docid   termid  freq    clabel
1       1       1       2
1       2       1       2
1       4       1       2
2       2       1       2
2       3       1       2
3       3       1       1
3       4       1       1
4       4       1       1
```

(Naive Bayes Classification) 문서 $d \in D$를 다음을 만족하는 부류 c^*로 분류한다.

$$c^* = \arg\max_c P(c) \prod_{v \in d} P(v|c)$$

여기서 $P(c)$는 특정 부류 c의 사전 확률이다. $P(v|c)$는 부류 c에 속하는 문서에서 단어 v가 나타날 확률이다.

(EM Clustering) clabel이 없는 데이터를 이용하고 2개의 cluster를 가정하여라. EM 알고리즘으로 MLE를 찾고 다음을 제시하여라.

- $\log P(c)$: prior probability of the cluster c.
- $\log P(v|c)$: conditaionl probability of the word v given the cluster c.
- $P(c|d)$: responsibilities.
- log-likelihood