**데이터 샘플**

@DOCUMENT

#DocID : 30821

#CAT'03: /건강과 의학/의약학/한의학 전통의학

#CAT'07: /정치/정부부처/보건복지부;/건강과 의학/한의학 전통의학

#TITLE : 96년부터 보건소에 한의사 공중보건의로 배치

#TEXT  :

 오는 96년부터 전국의 보건소에 한의사가 공중보건의로 배치돼 농어민들에

게 한방의료서비스가 본격 제공된다.

 보사부는 5일 병역법이 작년말 개정돼 한의사도 의사와 치과의사처럼 공중

보건의로 임용할 수 있게 됨에 따라 한방진료수요가 많은 농어촌 소재 보건

소에 우선적으로 공중보건의를 배치키로 했다.

<KW>

개정

계획

공중보건의

교육

농어민

 ---------------------------------------------------------------------------------------------------

* 위 예처럼 코퍼스의 각 문서에서 어절(feature)은 TEXT 태그 아래에 있는 것만 사용
* 카테고리는 #CAT'03 의 첫 항목 예) #CAT'03: /건강과 의학
* <KW> 아래에 있는 단어들은 특징에서 제외.
* 반복 어절 제외
* 구두점 제외
* '-' , ',' , '/' , ' .' , '<' , '>', ' " ' , ' [' , ']' , ' {', '}' 제외하고 공백 삽입, 예) 초등학교-중학교 --> 초등학교 중학교 (각각 어절로서 feature)
* 카테고리 번호

0 건강과 의학

1 경제

2 과학

3 교육

4 문화와 종교

5 사회

6 산업

7 여가생활

* feature 개수가 다소 많아 PC(노트북) 수준에서 메모리 문제를 확인해 보지는 않았지만 실제 차지 하는 용량은 많지 않아 문제 없을 것으로 것으로 생각됨.
* 만약 문제가 생기는 경우 교수에게 보고하고 데이터 량을 축소 (파일 1~3을 train, 4을 test 사용)

**학습데이터 파일 예시 (테스트 파일도 동일)**

XXX.train

(반드시 인덱스 오름차순, 카이스퀘어 사용 [average 와 max중 max사용], 한줄마다 각 파일 하나임)

4 0:0.0016025641 1:0.08125 2:0.006325911 3:0.08125 4:0.08125

6 0:0.0016025641 2:0.006325911 5:0.03076923 6:0.03076923 7:0.006325911 8:0.03076923 9:0.03076923 10:0.03076923 11:0.0016025641

6 0:0.0016025641 11:0.0016025641 12:0.03076923 13:0.03076923

4 7:0.006325911 11:0.0016025641 14:0.08125

첫 줄에서 왼쪽부터

4 🡪

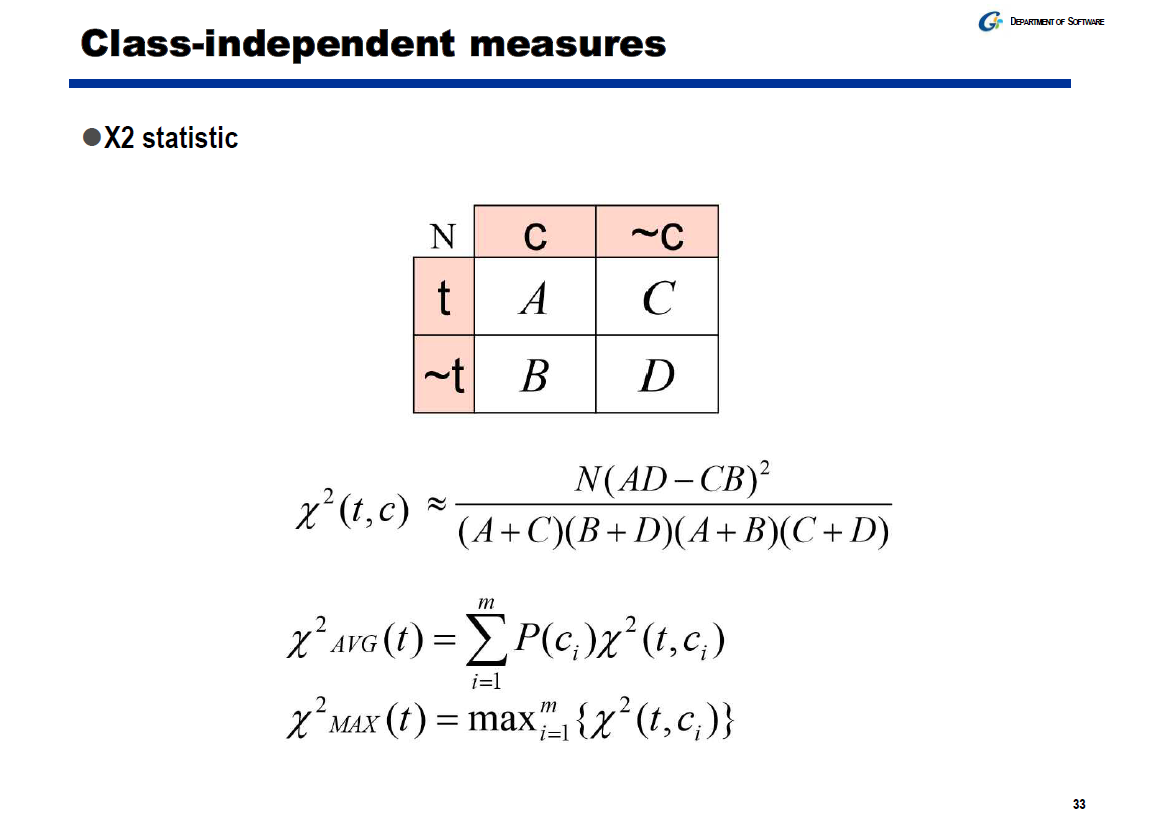
* + - * + 카테고리 번호

0:0.0016025641 🡪

* + - * + 콜론 왼쪽(0)은 문서에서 나타난 feature의 인덱스 번호 (feature 인덱스는 카이제곱 통계량이 큰 순서대로 0부터 할당, 예시는 카이제곱 통계량 순서대로 작성되어 있지 않음)
        + 문서에서 나타난 feature의 인덱스는 왼쪽부터 오름차순으로 작성해야 함
        + 콜론 오른쪽(0.0016025641)은 가중치 (카이제곱 통계량을 사용)
        + :~나오는 값이 순서가 아닌건 이 예시가 정리를 안한 것

1:0.08125 🡪

* + - * + 위와 동일
        + feature의 인덱스가 1이므로 0보다 오른쪽에 작성

++++

C – 카테고리 1개

~C – 나머지 카테고리

T – 어절

MAX 사용

Feture 중 값이 큰 순서로 index를 0부터 부여서 해서 나열

