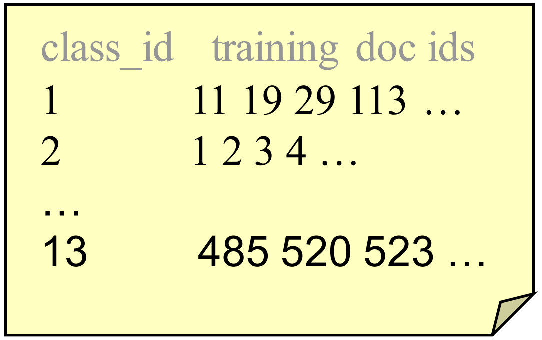
Programming Assignment 3 (1/3)

- **Multinomial NB Classifier**:
  - Text collection:
    - The 1095 news documents.
    - 13 classes (id 1~13), each class has 15 training documents.
      - https://ceiba.ntu.edu.tw/course/88ca22/content/training.txt



class_id	training doc ids
1	11 19 29 113 …
2	1 2 3 4 …
…	
13	485 520 523 …

training.txt

- The remaining documents are for testing.
  - Send your result to Kaggle.
  - See kaggle教學詳細版.pdf for the detail of the output format

Programming Assignment 3 (2/3)

- Note:
  - For each class, you have to calculate *M* $P(X=t|c)$ parameters.
    - *M* is the size of your vocabulary.
  - Then, the total number of parameters in your system will be $|C|*M$ ← can be a huge number.
  - We know that many terms in the vocabulary are not indicative.
  - **Employ at least one feature selection method** and use only **500 terms** in your classification.
    - X^2 test.
    - Likelihood ratio.
    - Pointwise/expected MI.
    - Frequency-based methods.
  - When classify a testing document, terms not in the selected vocabulary are ignored.

Programming Assignment 3 (3/3)

- To avoid zero probabilities, calculate *P*(*X*=*t*|*c*) by using add-one smoothing.

$$P(X = t_k \mid c) = \frac{T_{ct_k} + 1}{{}_{t' \;\; V}(T_{ct'} + 1)} = \frac{T_{ct_k} + 1}{{}_{t' \;\; V}(T_{ct'}) + |V|}$$

- Test your result on Kaggle !!
  - https://www.kaggle.com/t/001ab107135541378752ca9215000af0

- Please zip and submit $^{1.}$source code and $^{2.}$a report to TA.
  - 3 weeks to complete, that is, **2022/1/4**.