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/b079e8/content/training.txt

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

training.txt

doc_id	class_id
7	2
14	8
22	11
23	11
…	

output.txt

    - The remaining documents are for testing.
      - Generate an output file (output.txt) that records your classification results.
      - Exclude all training documents.
      - Ascending order to doc_id.

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 a 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.

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

- Please zip and submit 1. your classification result (output.txt), 2. source code, and 3. a report to TA.
  - 3 weeks to complete, that is, **2012/12/18**.

- TA will announce best micro/macro-averaging precision, recall, and F1.