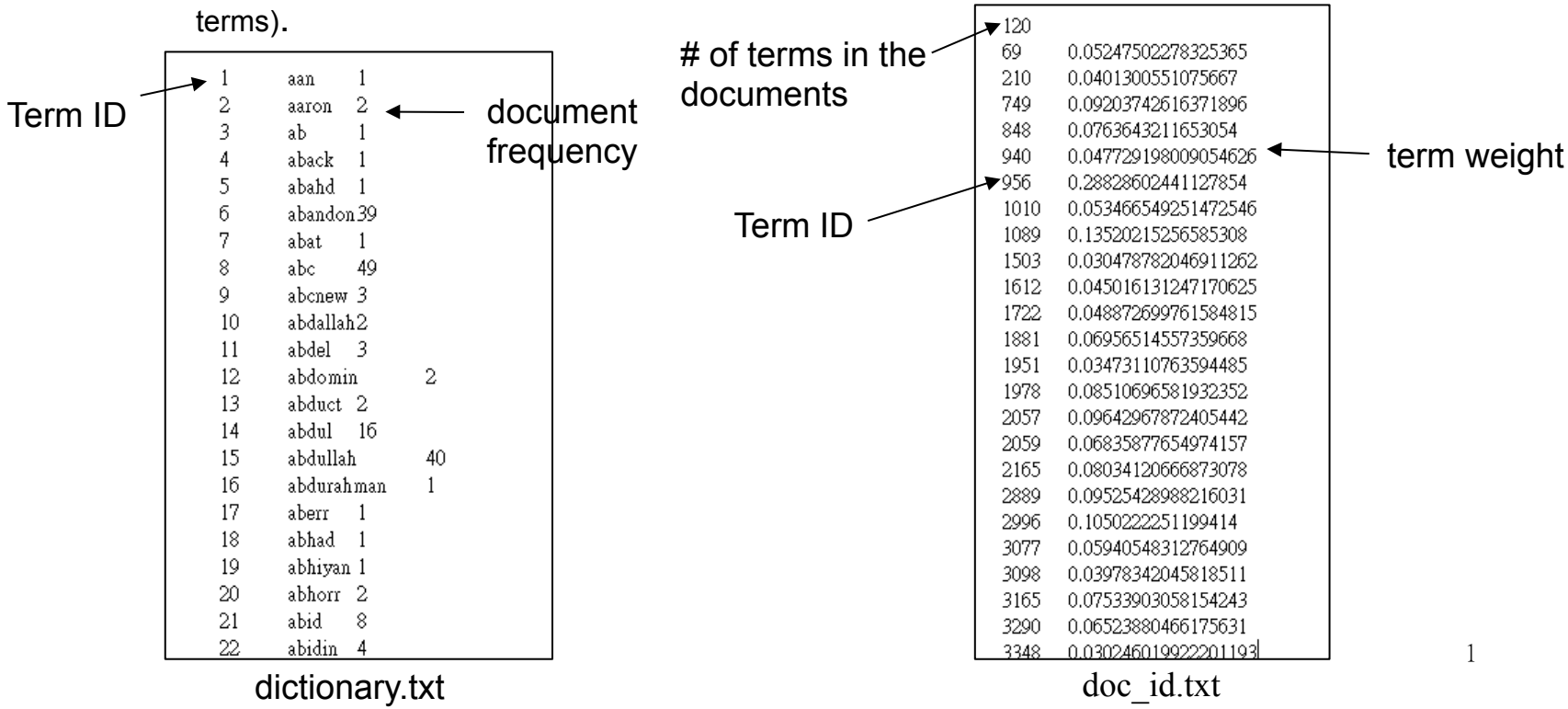
Programming Assignment (1/3)

- **Rocchio Classifier**:
  - Text collection (https://ceiba.ntu.edu.tw/course/99b512/content/PA.zip)
    - The 1095 news documents.
    - Each document is represented as high dimensional term vector (12299 terms).



Programming Assignment (2/3)

- Training dataset
  - 13 classes (id 1~13), each class has 15 training documents.

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.
    - **use Euclidean distance as the underlying distance measure!!**

$$\left|\underline{x}-\underline{y}\right| = \sqrt{\sum_{i=1}^{M}(x_i - y_i)^2}$$

    - Exclude all training documents.
    - Ascending order to doc_id.

Programming Assignment 3 (3/3)

- 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, **2013/12/11**.

- TA will check your micro/macro-averaging precision, recall, and F1.