

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).

Term ID →

| | | |
|----|------------|----|
| 1 | aan | 1 |
| 2 | aaron | 2 |
| 3 | ab | 1 |
| 4 | aback | 1 |
| 5 | abahd | 1 |
| 6 | abandon | 39 |
| 7 | abat | 1 |
| 8 | abc | 49 |
| 9 | abcnew | 3 |
| 10 | abdallah | 2 |
| 11 | abdel | 3 |
| 12 | abdomin | 2 |
| 13 | abduct | 16 |
| 14 | abdul | 40 |
| 15 | abdullah | 1 |
| 16 | abdurahman | 1 |
| 17 | aberr | 1 |
| 18 | abhad | 1 |
| 19 | abhiyan | 1 |
| 20 | abhorr | 2 |
| 21 | abid | 8 |
| 22 | abidin | 4 |

document frequency

dictionary.txt

of terms in the documents

Term ID →

| | |
|------|----------------------|
| 120 | 0.05247502278325365 |
| 69 | 0.0401300551075667 |
| 210 | 0.09203742616371896 |
| 749 | 0.0763643211653054 |
| 848 | 0.047729198009054626 |
| 940 | 0.28828602441127854 |
| 956 | 0.053466549251472546 |
| 1010 | 0.13520215256585308 |
| 1089 | 0.030478782046911262 |
| 1503 | 0.045016131247170625 |
| 1612 | 0.048872699761584815 |
| 1722 | 0.06956514557359668 |
| 1881 | 0.03473110763594485 |
| 1951 | 0.08510696581932352 |
| 1978 | 0.09642967872405442 |
| 2057 | 0.06835877654974157 |
| 2059 | 0.08034120666873078 |
| 2165 | 0.09525428988216031 |
| 2889 | 0.1050222251199414 |
| 2996 | 0.05940548312764909 |
| 3077 | 0.03978342045818511 |
| 3098 | 0.07533903058154243 |
| 3165 | 0.06523880466175631 |
| 3290 | 0.030246019922201193 |
| 3348 | |

term weight

doc_id.txt

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!!**

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

- 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.