We have 10 documents as follows:

|  |  |
| --- | --- |
| **No.** | **Contents** |
| D1. | AABAAABBBAACD |
| D2. | ACDAAABBBCCCDDD |
| D3. | CCDDDCDDDDDCCC |
| D4. | ABCDEABCDEABCDE |
| D5. | ABCCCCCCCCCCBC |
| D6. | DEADEADDDDDE |
| D7. | BEBBBBEEECBBBC |
| D8. | ABCAAAABBE |
| D9. | AEEEEABCEEA |
| D10. | ABCBCBCBCDDD |

Assume that we want to find the documents satissfying the following conditions.

1. Having A and B
2. At least 3 A’s or 3 B’s
3. The difference between #A and #B is not greater than 2

In order to do so, we firstly remove documents that violating the first condition. Then, we construct a *distance function* whose parameters are the remaining documents, which yield the following results.

d(D1) 0.053

d(D2) 0.228

d(D4) 0.377

d(D5) 0.127

d(D8) 0.072

d(D9) 0.315

d(D10) 0.188

We then set a *threshold T* to predict the documents. Documents whose distance function value is less or equal to T will be marked a positive and vice versa. With *T* = 0.2

1. Construct the confusion matrix
2. Calculate the value of accuracy, recall, precision, F-measure
3. Which value of *T* will give the best results for the above metrics