HW4

Qingyuan Zhang

Results:

=========

1: 1.1363 [593,241,353,402,303,518,269]

1: 1.2636 [946,269,270,116,319,381,378]

1: 0.5462 [904,137,396,109,437,374,322]

1: 0.3587 [760,83,518,108,521,374,315]

1: 0.2014 [706,80,523,108,574,374,314]

1: 0.1040 [683,78,537,108,586,374,313]

1: 0.0605 [661,78,558,108,586,375,313]

1: 0.0462 [628,78,589,108,586,375,315]

1: 0.0506 [589,78,628,108,588,375,313]

1: 0.0624 [538,78,679,108,589,375,312]

1: 0.0783 [490,78,731,108,586,374,312]

1: 0.0687 [459,78,765,108,583,374,312]

1: 0.0498 [450,78,775,108,583,374,311]

1: 0.0345 [444,78,779,108,585,374,311]

1: 0.0231 [442,78,784,108,584,373,310]

1: 0.0188 [442,78,784,108,584,373,310]

1: 0.0105 [441,78,785,108,584,373,310]

1: 0.0111 [440,78,787,108,583,373,310]

1: 0.0128 [437,78,790,108,583,373,310]

1: 0.0058 [436,78,791,108,583,373,310]

1: 0.0065 [435,78,792,108,583,373,310]

1: 0.0074 [434,78,793,108,583,373,310]

1: 0.0000 [434,78,793,108,583,373,310]

NW: 69.82 (303/434)

We: 98.72 (77/78)

Am: 91.93 (729/793)

BC: 81.48 (88/108)

BC: 43.91 (256/583)

CT: 97.59 (364/373)

BN: 26.13 (81/310)

Score: 0.7084733109369168

The cosine similarity has a good performance in We, Am and CT. However, for BN and BC the cosine similarity does not perform very well. Overall, it generate a better result than Euclidean distance at separating the documents into equal-sized cluster.

However, it is not true that cosine similarity is always better than Euclidean. Our dataset is more easily to allow cosine similarity to separate but if we change to some other dataset, Euclidean may over-performance.