Hw4 | Allen Herrera | CS325

Task

- ☑ Override distance in KmeanHerrea
- ☑ Submit code for cosine distance
- ☑ Describe observations about clusters

K-means-Herrera (Cosine)

```
   1: 6.8989 [593,241,353,402,303,518,269]
   1: 13.0126[946,269,270,116,319,381,378]
   1: 4.4983 [904,137,396,109,437,374,322]
   1: 4.0619 [760,83,518,108,521,374,315]
   1: 1.3751 [706,80,523,108,574,374,314]
   1: 0.9815 [683,78,537,108,586,374,313]
   1: 0.3549 [661,78,558,108,586,375,313]
   1: 0.2547 [628,78,589,108,586,375,315]
   1: 0.3248 [589,78,628,108,588,375,313]
   1: 0.4043 [538,78,679,108,589,375,312]
   1: 0.5161 [490,78,731,108,586,374,312]
   1: 0.4223 [459,78,765,108,583,374,312]
   1: 0.2896 [450,78,775,108,583,374,311]
   1: 0.1634 [444,78,779,108,585,374,311]
   1: 0.1285 [442,78,784,108,584,373,310]
   1: 0.0892 [442,78,784,108,584,373,310]
   1: 0.0498 [441,78,785,108,584,373,310]
   1: 0.0519 [440,78,787,108,583,373,310]
   1: 0.0625 [437,78,790,108,583,373,310]
   1: 0.0282 [436,78,791,108,583,373,310]
   1: 0.0310 [435,78,792,108,583,373,310]
   1: 0.0358 [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
```

K-means-Euclidean

```
   1: 11.4285 [718,227,611,264,378,295,186]
   1: 13.9987 [299,286,920,109,263,165,637]
   1: 4.0345 [243,362,972,103,250,120,629]
   1: 2.1984 [214,392,1006,100,258,104,605]
   1: 1.3636 [198,404,1032,100,266,95,584]
   1: 1.2748 [193,410,1049,100,274,87,566]
   1: 1.3226 [184,415,1064,100,280,81,555]
   1: 1.0936 [175,419,1076,100,286,75,548]
   1: 0.9626 [169,420,1086,100,290,71,543]
   1: 0.5885 [170,420,1089,100,291,70,539]
   1: 0.5133 [169,420,1090,100,293,68,539]
   1: 0.3849 [167,422,1090,100,294,67,539]
   1: 0.2564 [164,423,1091,100,294,67,540]
   1: 0.0000 [164,423,1091,100,294,67,540]
NW: 83.54 (137/164)
BC: 47.28 (200/423)
Am: 72.14 (787/1091)
BC: 82.00 (82/100)
CT: 100.00 (294/294)
CT: 100.00 (67/67)
NW: 40.19 (217/540)
Score: 0.665920119447555
```

The first thing to notice is that the Cosine distance better clusters the documents. KmeansHerrera achieves 70% correct cluster categorization, while Euclidean achieves 66%. The next thing to look at is the first iteration as the 1st centroids are most influential to the clustering algorithm. Here we notice that the documents are more evenly distributed in the Cosine version than in the Euclidean distance version. This opens up opportunities for the documents to fall into the right cluster as each cluster is well represented and does grow or shrink too quickly. For example notice how in my Herrera version, the third cluster (Am) starts off with 353 then grows to 793 documents. In that cluster it maintains 91.93% accuracy, while the Euclidean version starts at 611, grows to 1091 documents and only captures an additional 58 documents correctly (72.14%) at the expense of getting an additional 240 documents clustered wrong.

Another thing to notice is how quickly the cluster (BC) converges in both tests. However because Euclidean creates three monstrous clusters with unimpressive accuracy, this boosts the % correct for other clusters as they have less documents to cluster. I.e (67/67) and (294/294) as opposed to 364/373 in KmeanHerrera.

A third observation from the more even distribution from the 1st iteration is that KmeanHerrera was actually able to guess the clusters (BN) and (We) whereas Euclidean couldn't do that. A small observation is that the Cosine version had more iterations of re-computing centroids as opposed to Euclidean.

Last interesting point, in KmeanHerrera, the first cluster (NW) grows to 946 documents in the 2nd adjustment but then shrinks from then on down to 434 documents with an accuracy of 69.82%.