CS5344 Big Data Analytics Technology
Lab 2 (AY2019/2020 Semester 2)
Analysis Report Task B

I tried to implement Kmean algorithm in Task B.

Here are results when I tested on the datafiles.zip (include 35files) with k from 2 to 8 for both Cosine Distance and Euclidean Distance:

- Iterations of the Kmean are same for both distances when it reached convergence
- The Kmean gives same items and number of items for each cluster corresponding in Cosine Distance and Euclidean Distance

I think my Kmean algorithm implement is not good enough to see what differences of performance and cluster numbers between using Cosine Distance and Euclidean Distance

1. **K = 2**

- **Cosine Distance:**

group_0 : [u'f1.txt', u'f8.txt', u'f12.txt', u'f13.txt', u'f11.txt', u'f14.txt']

group_1 : [u'f19.txt', u'f3.txt', u'f30.txt', u'f25.txt', u'f16.txt', u'f6.txt', u'f7.txt', u'f32.txt', u'f34.txt', u'f10.txt', u'f35.txt', u'f28.txt', u'f15.txt', u'f22.txt', u'f18.txt', u'f20.txt', u'f33.txt', u'f23.txt', u'f17.txt', u'f5.txt', u'f4.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f26.txt', u'f9.txt', u'f2.txt', u'f31.txt']

- **Euclidean Distance:**

group_0 : [u'f1.txt', u'f8.txt', u'f12.txt', u'f13.txt', u'f11.txt', u'f14.txt']

group_1 : [u'f19.txt', u'f3.txt', u'f30.txt', u'f25.txt', u'f16.txt', u'f6.txt', u'f7.txt', u'f32.txt', u'f34.txt', u'f10.txt', u'f35.txt', u'f28.txt', u'f15.txt', u'f22.txt', u'f18.txt', u'f20.txt', u'f33.txt', u'f23.txt', u'f17.txt', u'f5.txt', u'f4.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f26.txt', u'f9.txt', u'f2.txt', u'f31.txt']

2. **K = 3**

- **Cosine Distance**

group_0 : [u'f1.txt', u'f3.txt', u'f6.txt', u'f10.txt', u'f15.txt', u'f5.txt', u'f13.txt', u'f11.txt']

group_1 : [u'f19.txt', u'f30.txt', u'f25.txt', u'f16.txt', u'f34.txt', u'f35.txt', u'f28.txt', u'f22.txt', u'f18.txt', u'f33.txt', u'f23.txt', u'f17.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f26.txt', u'f31.txt']

group_2 : [u'f8.txt', u'f7.txt', u'f32.txt', u'f12.txt', u'f20.txt', u'f4.txt', u'f9.txt', u'f2.txt', u'f14.txt']

- **Euclidean Distance**

group_0 : [u'f1.txt', u'f3.txt', u'f6.txt', u'f10.txt', u'f15.txt', u'f5.txt', u'f13.txt', u'f11.txt']

group_1 : [u'f19.txt', u'f30.txt', u'f25.txt', u'f16.txt', u'f34.txt', u'f35.txt', u'f28.txt', u'f22.txt', u'f18.txt', u'f33.txt', u'f23.txt', u'f17.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f26.txt', u'f31.txt']

group_2 : [u'f8.txt', u'f7.txt', u'f32.txt', u'f12.txt', u'f20.txt', u'f4.txt', u'f9.txt', u'f2.txt', u'f14.txt']

3. **K = 4**

- **Cosine Distance**

group_0 : [u'f33.txt', u'f23.txt', u'f11.txt']

group_1 : [u'f16.txt', u'f32.txt', u'f28.txt', u'f15.txt', u'f26.txt']

group_2 : [u'f19.txt', u'f30.txt', u'f25.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f20.txt', u'f17.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f31.txt']

group_3 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f10.txt', u'f5.txt', u'f4.txt', u'f13.txt', u'f9.txt', u'f2.txt', u'f14.txt']

- **Euclidean Distance**

group_0 : [u'f33.txt', u'f23.txt', u'f11.txt']

group_1 : [u'f16.txt', u'f32.txt', u'f28.txt', u'f15.txt', u'f26.txt']

group_2 : [u'f19.txt', u'f30.txt', u'f25.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f20.txt', u'f17.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f31.txt']

group_3 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f10.txt', u'f5.txt', u'f4.txt', u'f13.txt', u'f9.txt', u'f2.txt', u'f14.txt']

4. **K =5**

- **Cosine Distance**

group_0 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f4.txt', u'f2.txt']

group_1 : [u'f33.txt', u'f23.txt', u'f11.txt']

group_2 : [u'f16.txt', u'f32.txt', u'f28.txt', u'f15.txt', u'f26.txt']

group_3 : [u'f19.txt', u'f30.txt', u'f25.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f20.txt', u'f17.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f31.txt']

group_4 : [u'f10.txt', u'f5.txt', u'f13.txt', u'f9.txt', u'f14.txt']

- **Euclidean Distance**

group_0 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f4.txt', u'f2.txt']

group_1 : [u'f33.txt', u'f23.txt', u'f11.txt']

group_2 : [u'f16.txt', u'f32.txt', u'f28.txt', u'f15.txt', u'f26.txt']

group_3 : [u'f19.txt', u'f30.txt', u'f25.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f20.txt', u'f17.txt', u'f21.txt', u'f27.txt', u'f29.txt', u'f24.txt', u'f31.txt']

group_4 : [u'f10.txt', u'f5.txt', u'f13.txt', u'f9.txt', u'f14.txt']

5. **K =6**

- **Cosine Distance**

group_0 : [u'f30.txt', u'f16.txt', u'f28.txt', u'f20.txt', u'f14.txt']

group_1 : [u'f25.txt', u'f23.txt', u'f27.txt', u'f24.txt']

group_2 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f10.txt', u'f5.txt', u'f4.txt', u'f9.txt', u'f2.txt']

group_3 : [u'f33.txt', u'f11.txt']

group_4 : [u'f32.txt', u'f15.txt', u'f13.txt', u'f26.txt']

group_5 : [u'f19.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f17.txt', u'f21.txt', u'f29.txt', u'f31.txt']

- **Euclidean Distance**

group_0 : [u'f30.txt', u'f16.txt', u'f28.txt', u'f20.txt', u'f14.txt']

group_1 : [u'f25.txt', u'f23.txt', u'f27.txt', u'f24.txt']

group_2 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f10.txt', u'f5.txt', u'f4.txt', u'f9.txt', u'f2.txt']

group_3 : [u'f33.txt', u'f11.txt']

group_4 : [u'f32.txt', u'f15.txt', u'f13.txt', u'f26.txt']

group_5 : [u'f19.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f17.txt', u'f21.txt', u'f29.txt', u'f31.txt']

6. **K =7**

- **Cosine Distance**

group_0 : [u'f30.txt', u'f16.txt', u'f28.txt', u'f20.txt']

group_1 : [u'f25.txt', u'f23.txt', u'f27.txt', u'f24.txt']

group_2 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f4.txt', u'f2.txt']

group_3 : [u'f33.txt', u'f11.txt']

group_4 : [u'f32.txt', u'f15.txt', u'f26.txt']

group_5 : [u'f19.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f17.txt', u'f21.txt', u'f29.txt', u'f31.txt']

group_6 : [u'f10.txt', u'f5.txt', u'f13.txt', u'f9.txt', u'f14.txt']

- **Euclidean Distance**

group_0 : [u'f30.txt', u'f16.txt', u'f28.txt', u'f20.txt']

group_1 : [u'f25.txt', u'f23.txt', u'f27.txt', u'f24.txt']

group_2 : [u'f1.txt', u'f8.txt', u'f3.txt', u'f6.txt', u'f7.txt', u'f4.txt', u'f2.txt']

group_3 : [u'f33.txt', u'f11.txt']

group_4 : [u'f32.txt', u'f15.txt', u'f26.txt']

group_5 : [u'f19.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f12.txt', u'f18.txt', u'f17.txt', u'f21.txt', u'f29.txt', u'f31.txt']

group_6 : [u'f10.txt', u'f5.txt', u'f13.txt', u'f9.txt', u'f14.txt']

7. **K = 8**

- **Cosine Distance**

group_0 : [u'f1.txt', u'f3.txt', u'f12.txt', u'f4.txt', u'f2.txt']

group_1 : [u'f30.txt', u'f16.txt', u'f28.txt', u'f20.txt']

group_2 : [u'f25.txt', u'f23.txt', u'f27.txt', u'f24.txt']

group_3 : [u'f8.txt', u'f6.txt', u'f7.txt']

group_4 : [u'f33.txt', u'f11.txt']

group_5 : [u'f32.txt', u'f15.txt', u'f26.txt']

group_6 : [u'f19.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f18.txt', u'f17.txt', u'f21.txt', u'f29.txt', u'f31.txt']

group_7 : [u'f10.txt', u'f5.txt', u'f13.txt', u'f9.txt', u'f14.txt']

- **Euclidean Distance**

group_0 : [u'f1.txt', u'f3.txt', u'f12.txt', u'f4.txt', u'f2.txt']

group_1 : [u'f30.txt', u'f16.txt', u'f28.txt', u'f20.txt']

group_2 : [u'f25.txt', u'f23.txt', u'f27.txt', u'f24.txt']

group_3 : [u'f8.txt', u'f6.txt', u'f7.txt']

group_4 : [u'f33.txt', u'f11.txt']

group_5 : [u'f32.txt', u'f15.txt', u'f26.txt']

group_6 : [u'f19.txt', u'f34.txt', u'f35.txt', u'f22.txt', u'f18.txt', u'f17.txt', u'f21.txt', u'f29.txt', u'f31.txt']

group_7 : [u'f10.txt', u'f5.txt', u'f13.txt', u'f9.txt', u'f14.txt']