Clustering text data with K-means

7 questions

1 point

1.
(True/False) The clustering objective (heterogeneity) is non-increasing for this example.

- True
- False

1 point

2.
Let's step back from this particular example. If the clustering objective (heterogeneity) would ever increase when running K-means, that would indicate: (choose one)

- K-means algorithm got stuck in a bad local minimum
- There is a bug in the K-means code
- All data points consist of exact duplicates
- Nothing is wrong. The objective should generally go down sooner or later.

1 point

3.

Refer to the output of K-means for K=3 and seed=0. Which of the three clusters contains the greatest number of data points in the end?

- Cluster #0
- Cluster #1
- Cluster #2

---

1 point

4.

Another way to capture the effect of changing initialization is to look at the distribution of cluster assignments. Compute the size (# of member data points) of clusters for each of the multiple runs of K-means.

Look at the size of the largest cluster (most # of member data points) across multiple runs, with seeds 0, 20000, ..., 120000. What is the **maximum** value this quantity takes?

18132

---

1 point

5.

Refer to the section "Visualize clusters of documents". Which of the 10 clusters above contains the **greatest** number of articles?

- Cluster 0: artists, poets, writers, environmentalists
- Cluster 4: track and field athletes
- Cluster 5: composers, songwriters, singers, music producers
- Cluster 7: baseball players
- Cluster 9: lawyers, judges, legal scholars

---

1 point

6.

Refer to the section "Visualize clusters of documents". Which of the 10 clusters above contains the **least** number of articles?

- Cluster 1: film directors
- Cluster 3: politicians
- Cluster 6: soccer (football) players
- Cluster 7: baseball players
- Cluster 9: lawyers, judges, legal scholars

1 point

7.

Another sign of too large K is having lots of small clusters. Look at the distribution of cluster sizes (by number of member data points). How many of the 100 clusters have fewer than 236 articles, i.e. 0.4% of the dataset?

35

Submit Quiz