K-means Clustering assignment

Andreas Kollias

Objective: Group 6 documents into 2 clusters based on their TF-IDF word scores.

We have collected 6 opinion articles on Trump's policies, published in a mainstream online news portal. We want to cluster together in two clusters opinion articles depending on the TF-IDF score of the words appearing in these articles.

Step 1: TF-IDF score (done)

First, we calculated the TF-IDF score for each word in each of the 6 opinion articles. In the table below are the TF-IDF scores of 9 words with the highest TF-IDF scores in these opinion articles. In each opinion article some words have a high score. The same words may have a low score in other articles in our collection. This is an indication that these opinion articles focus on different issues regarding Trump's policies.

Docs	America First	Defend Democracy	Deregulation	Resist	Tax Cuts	Tax the Rich	Anti-Authori-tarianism	Travel Bans	Anti-"Woke"
Doc1	0.44	0.002	0.32	0.09	0.22	0.1	0.02	0.24	0.13
Doc2	0.01	0.32	0.03	0.27	0.02	0.22	0.29	0.02	0.023
Doc3	0.09	0.26	0.06	0.22	0.013	0.29	0.321	0.03	0.003
Doc4	0.34	0.03	0.319	0.003	0.26	0.07	0.021	0.28	0.29
Doc5	0.29	0.002	0.35	0.004	0.17	0.007	0.09	0.44	0.3
Doc6	0.0023	0.29	0.004	0.39	0.034	0.22	0.32	0.045	0.006

Step-by-Step Guide (steps you must do)

2. Choose k = 2

You want 2 clusters, so you set k = 2.

3. Initialize Centroids

Choose which 2 docs will be the first 2 clusters (remember k=2) (In this assignment we "randomly" chose Doc4 and Doc6). Their centroids are the TF-IDF scores in each of the most important words in the collection of 6 documents.

- Centroid 1 = Doc4 vector
- Centroid 2 = Doc6 vector

4. Compute Distance to Centroids

- For each document, compute its Euclidean distance to each centroid:

$$\text{Distance}(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

- Assign each document to a cluster.

5. Recalculate Centroids

- For each cluster, you collect all the documents assigned to it (based on closest distance), and then:

Document	peace	negotiations
Doc1	0.32	0.26
Doc3	0.34	0.21
Doc4	0.20	0.33

- compute the average TF-IDF for each column (each of the 9 terms in the assignment). This average becomes the new centroid for the next iteration.

6. For each document

- compute its Euclidean distance to each new cluster centroid
- Reassign documents based on new distances.

For more iterations. Repeat Steps 4–6 (skip this)

- Reassign documents based on new distances.
- Recalculate centroids again.
- Continue iterating until:
    - Assignments no longer change (convergence)
    - Or a maximum number of iterations is reached

7. Final Output

After convergence, you'll have:

- 2 clusters, each containing some subset of the documents.
- Each document assigned to one cluster.

8. Conclusions