
Clustering Wikipedia Articles

Lane Aasen

Department of Computer Science
University of Washington
Seattle, WA 98105
aaasen@cs.washington.edu

Abstract

Clustering Wikipedia articles using unsupervised learning techniques including K-Means, Latent Dirichlet Allocation (LDA), and spectral clustering.

1 Dataset

The provided dataset contains 15,903 Wikipedia articles in tf-idf format. There are 10,574 unique words in this dataset. Each document is represented as a sparse vector with one dimension for each word. I have not experimented with dimensionality reduction yet, but it could be an interesting topic to explore.

2 K-Means Clustering

For the project milestone, I have implemented K-Means clustering on the provided subset of Wikipedia articles.

2.1 Choosing K

2.1.1 Minimizing Distortion

Given K clusters C_1, C_2, \dots, C_K where each cluster is a set of document vectors and μ_i is the centroid of C_i , the total distortion is defined as follows:

$$\sum_{i=1}^K \sum_{d \in C_i} \|d - \mu_i\|^2$$

To minimize the distortion, we could set K equal to the number of documents, but then the clusters would be meaningless. We want to choose a K with low distortion that also results in interpretable clusters. Figure 1 shows a plot of K versus total distortion. When $1 \leq K \leq 16$, adding additional clusters has a large impact on the distortion. However, once $K > 16$, adding additional clusters has little impact on the distortion. From this alone, it makes sense to set $K = 16$ since it provides a good balance of distortion and interpretability.

2.2 K and Cluster Size

As K increases, the clusters become more sparse. Once $K = 256$, over half of the clusters have only one document, and are essentially useless. When $K = 16$, the median cluster size is 8.5, and the cluster sizes are as follows:

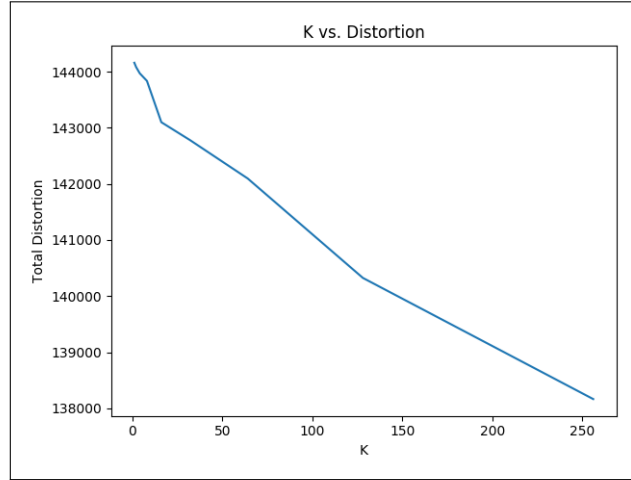


Figure 1: K versus total distortion for $K \in \{1, 2, 4, \dots, 256\}$

[10061, 3013, 1128, 909, 707, 30, 23, 13, 4, 4, 3, 2, 2, 2, 1, 1]

Over half of the clusters are very small, and one of the clusters is too large to be interpretable. This indicates that the data has significant outliers and lacks a structure conducive to clustering.

In a perfect dataset with n documents and K clusters, we would expect each cluster to have exactly K/n documents.

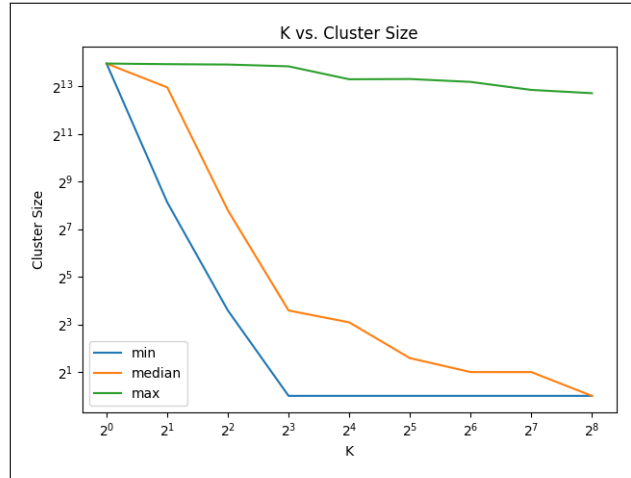


Figure 2: K versus minimum, median, and maximum cluster size for $K \in \{1, 2, 4, \dots, 256\}$ with a \log_2 scale on both axes.

2.3 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Table 1: Sample table title

| CLUSTER | SIZE |
|----------------|-------------|
|----------------|-------------|