**Compile and Run Instructions**:

**Commands to run**:

The argument ‘numberOfClusters’ is optional.

1. If given, use following to execute the file.

TweetsClusteringUsingKMeans.py <numberOfClusters> <initialSeedsFile> <TweetsDataFile> <outputFile>

1. Else, then default value for ‘numberOfClusters’ will be taken as 25

TweetsClusteringUsingKMeans.py <initialSeedsFile> <TweetsDataFile> <outputFile>

**Example commands**:

1. python TweetsClusteringUsingKMeans.py ﻿25 InitialSeeds.txt Tweets.json tweets-k-means-output.txt
2. python TweetsClusteringUsingKMeans.py ﻿InitialSeeds.txt Tweets.json tweets-k-means-output.txt

**Results:**

After running TweetsClusteringUsingKMeans.py , a file tweets-k-means-output.txt is created in your current directory which contains the results after clustering the tweets based on the initialSeeds.txt containing k initial selected cluster centroids.

**Values of SSE for different number of clusters:**

|  |  |
| --- | --- |
| **numberOfClusters** | **SSE** |
| 5 | 28.24255215943798 |
| 10 | 23.11550935670716 |
| 15 | 16.26499456060448 |
| 20 | 13.317684028435771 |
| 25 | 8.833204597723492 |

**Example output file:**

An example output file for numberOfClusters=25 is present there in folder as ‘tweets-k-means-output.txt’.

If you want to change the number of clusters, you will have to change number of seeds in ‘InitialSeeds.txt’ according to number of clusters you want. numberOfClusters=number of entries in the ‘InitialSeeds.txt’.