SathyaNarayanan Srinivasan

Sxs142031@utdallas.edu

Abstract

Perform K –Means using Euclidean and Jaccard Metrics

clustering

Machine Learning Assignment - 5

Table of Contents

[Part - 1 2](#_Toc435958287)

[Requirements 2](#_Toc435958288)

[Compiling and Running Instructions 2](#_Toc435958289)

[EXAMPLE COMMANDS 3](#_Toc435958290)

[**Output** 3](#_Toc435958291)

[Part - 2 8](#_Toc435958292)

[Requirements 8](#_Toc435958293)

[Compiling and Running Instructions 8](#_Toc435958294)

[EXAMPLES 9](#_Toc435958295)

[**Output** 9](#_Toc435958296)

# Part - 1

# Requirements

* **Operating System** – Windows 7 or Higher
* **Java:** [1.7.0\_45; Java HotSpot(TM) 64-Bit Server VM 24.45-b08](https://docs.oracle.com/javase/8/docs/technotes/guides/install/windows_jdk_install.html#CHDEBCCJ)
* **Runtime:** [Java(TM) SE Runtime Environment 1.7.0\_45-b18](https://docs.oracle.com/javase/8/docs/technotes/guides/install/windows_jre_install.html#CHDEDHAJ)
* **Notepad ++ / Sublime Text / Vi** – For reading the output

# Compiling and Running Instructions

* In Elevated Windows Command Prompt:
  + Navigate to the downloaded and unzipped project folder using cd command
  + Navigate to the Part1 folder using cd command
  + **TO COMPILE :**

USE: **javac MainClass.java**

* + **TO RUN**:

USE: **java MainClass <K> <ABSOLUTE PATH TO DATA FILE> <ABSOLUTE PATH TO OUTPUT FILE>**

* + **Some errors as unable to find the file specified OR the inbuilt error in the program might occur. This might be due to incorrect path specification. Please give the ABSOLUTE PATH to the data set. The folder structure might contain folders with spaces. Please enclose the entire path with “ ” in a Windows Machine. Please give the ABSOLUTE PATH correctly with respect to the OS.**
  + **Note that the k-means algorithm can converge to a local minimum, and also exhibit degeneracy, whereby one of the clusters has no members. Should these scenarios occur, simply re-run the algorithm**
  + **Please refer to the samples below:**

# EXAMPLE COMMANDS

**COMPILING**

*D:\Assignment 5\Part1>javac MainClass.java*

**RUNNING**

*D:\Assignment 5\Part1>java MainClass 25 test\_data.txt output.txt*

# **Output**

**SSE VALUES**

**K = 25**

|  |  |
| --- | --- |
| 1 | 0.01782472 |
| 2 | 0.02526824 |
| 3 | 0.03105182 |
| 4 | 0.04025849 |
| 5 | 0.05332383 |
| 6 | 0.05553117 |
| 7 | 0.26750717 |
| 8 | 0.29349432 |
| 9 | 0.30124632 |
| 10 | 0.32076632 |
| 11 | 1.07917632 |
| 12 | 1.08470237 |
| 13 | 1.09143741 |
| 14 | 1.10153741 |
| 15 | 1.10567929 |
| 16 | 1.18603569 |
| 17 | 1.19375074 |
| 18 | 1.21252934 |
| 19 | 1.23650001 |
| 20 | 1.24260626 |
| 21 | 1.24516876 |
| 22 | 1.24534526 |
| 23 | 1.2630346 |
| 24 | 1.27884885 |
| 25 | 1.29615829 |
| **Total** | **19.268783** |

**K = 6**

|  |  |
| --- | --- |
| 1 | 0.65848514 |
| 2 | 0.6970265 |
| 3 | 0.80301796 |
| 4 | 1.1819357 |
| 5 | 1.2363975 |
| 6 | 1.34617255 |
| **Total** | **5.92303535** |

**K = 10**

|  |  |
| --- | --- |
| 1 | 0.0896378 |
| 2 | 0.13831856 |
| 3 | 0.15872094 |
| 4 | 0.39222904 |
| 5 | 0.77416686 |
| 6 | 0.7918562 |
| 7 | 0.8410096 |
| 8 | 0.89365834 |
| 9 | 0.89961684 |
| 10 | 0.98161352 |
| **Total** | **5.9608277** |

**K= 15**

|  |  |
| --- | --- |
| 1 | 0.056815 |
| 2 | 0.09929168 |
| 3 | 0.1340481 |
| 4 | 0.15140897 |
| 5 | 0.16072897 |
| 6 | 0.20937107 |
| 7 | 0.21869907 |
| 8 | 0.23038907 |
| 9 | 0.28326813 |
| 10 | 0.38174626 |
| 11 | 0.38380326 |
| 12 | 0.38795226 |
| 13 | 0.4056416 |
| 14 | 0.4465176 |
| 15 | 0.47250475 |
| **Total** | **4.02218579** |

**K = 20**

|  |  |
| --- | --- |
| 1 | 0.00907896 |
| 2 | 0.02252823 |
| 3 | 0.04474571 |
| 4 | 0.07628101 |
| 5 | 0.11843249 |
| 6 | 0.14061649 |
| 7 | 0.16981614 |
| 8 | 0.17158948 |
| 9 | 0.57805148 |
| 10 | 1.25075848 |
| 11 | 1.25737948 |
| 12 | 1.27597348 |
| 13 | 1.34246973 |
| 14 | 1.3886344 |
| 15 | 1.40916689 |
| 16 | 1.41152408 |
| 17 | 1.41321908 |
| 18 | 1.44962968 |
| 19 | 1.4999232 |
| 20 | 1.5050292 |
| **Total** | **16.53484769** |

# 

# Part - 2

# Requirements

* **Operating System** – Windows 7 or Higher
* **Java:** [1.7.0\_45; Java HotSpot(TM) 64-Bit Server VM 24.45-b08](https://docs.oracle.com/javase/8/docs/technotes/guides/install/windows_jdk_install.html#CHDEBCCJ)
* **Runtime:** [Java(TM) SE Runtime Environment 1.7.0\_45-b18](https://docs.oracle.com/javase/8/docs/technotes/guides/install/windows_jre_install.html#CHDEDHAJ)
* **Notepad ++ / Sublime Text / Vi** – For reading the output

# Compiling and Running Instructions

* In Elevated Windows Command Prompt:
  + Navigate to the downloaded and unzipped project folder using cd command
  + Navigate to the Part 2 folder using cd command
  + **TO COMPILE :**

For Windows:

USE: **javac –cp *".;java-json.jar" KMeansTweetClustering.java***

For Unix OS:

USE: **javac –cp *".:java-json.jar" KMeansTweetClustering.java***

* + **TO RUN**:

For Windows:

USE: **java -cp .;java-json.jar KMeansTweetClustering <25> <ABSOLUTE PATH TO INITIAL SEEDS FILE> <ABSOLUTE PATH TO JSON FILE> <ABSOLUTE PATH TO OUTPUT FILE>**

For Unix :

USE: **java -cp .:java-json.jar KMeansTweetClustering <25> <ABSOLUTE PATH TO INITIAL SEEDS FILE> <ABSOLUTE PATH TO JSON FILE> <ABSOLUTE PATH TO OUTPUT FILE>**

* + **Some errors as unable to find the file specified OR the inbuilt error in the program might occur. This might be due to incorrect path specification. Please give the ABSOLUTE PATH to the data sets. The folder structure might contain folders with spaces. Please enclose the entire path with “ ” in a Windows Machine. Please give the ABSOLUTE PATH correctly with respect to the OS.**

# EXAMPLES

**COMPILING**

*D:\Assignment 5\Part2>javac -cp ".;java-json.jar" KMeansTweetClustering.java*

**RUNNING**

D:\DriveFiles\Fall15\ML\Assignment Solutions\Assignment 5\Part2>java -cp .;java-json.jar KMeansTweetClustering 25 InitialSeeds.txt Tweets.json output.txt

# **Output**

**SSE VALUES**

**K=25**

|  |  |
| --- | --- |
| 1 | 9.33730434 |
| 2 | 0.030276 |
| 3 | 0.52690248 |
| 4 | 0 |
| 5 | 0.10333058 |
| 6 | 17.38493164 |
| 7 | 0 |
| 8 | 0.02468642 |
| 9 | 1.12542734 |
| 10 | 0.80231928 |
| 11 | 2.2390605 |
| 12 | 0.02765952 |
| 13 | 0 |
| 14 | 10.10810086 |
| 15 | 0 |
| 16 | 0 |
| 17 | 0.54609736 |
| 18 | 0 |
| 19 | 0.5071138 |
| 20 | 2.0972462 |
| 21 | 0 |
| 22 | 1.60842506 |
| 23 | 0.17230888 |
| 24 | 10.226725 |
| 25 | 0.6486845 |
| **Total** | **57.51659976** |

**Cluster Values**

**<cluster-id> := <List of tweet ids separated by comma>**

0 := 323906398230544385,323906397962121216,323908795396943872,323907087551836160,323906399149109248,323906397735641088,323906398352195585,323906398012461057,323906397609791488,323906398314438656,323906656318676993,323906399300100096,323906397618196483,323906397853073410,323906398826164225,323906398993932289,323906399295926273,323907771256938496,323908795254312962,323908455545049088,323906398230544385,323906397962121216,323908795396943872,323907087551836160,323906399149109248,323906397735641088,323906398352195585,323906398012461057,323906397609791488,323906398314438656,323906656318676993,323906399300100096,323906397618196483,323906397853073410,323906398826164225,323906398993932289,323906399295926273,323907771256938496,323908795254312962,323908455545049088,

1 := 323916051614138368,323906485249789952,323915000567697409,323920146454425600,323923559799996417,323921510282702848,323911610236293120,323925264352546816,323906483584655360,323916051614138368,323906485249789952,323915000567697409,323920146454425600,323923559799996417,323921510282702848,323911610236293120,323925264352546816,323906483584655360,

2 := 323906653651079168,323906651209994241,323906650987692034,323906657333682176,323906653651079168,323906651209994241,323906650987692034,323906657333682176,

3 := 323963901769297921,323906398176030720,323932094190439874,323910330457669633,323907258301939713,323955716392112128,323906567294562306,323910330315075584,324226045052071936,323963901769297921,323906398176030720,323932094190439874,323910330457669633,323907258301939713,323955716392112128,323906567294562306,323910330315075584,324226045052071936,

4 := 324229792834674689,323909308188344320,324229792834674689,323909308188344320,

5 := 323906482489921537,323916814490939393,323922279182520320,323915624835342337,323950257832603648,323924665074581504,323913403460636673,323925432548335616,323917413789868034,323944113764438016,323932515599515649,323924496144797697,323930212121317377,323918009901150208,323929272685318147,323911097818173440,323927822072684544,323924155391164416,323913403951357953,323919716039155712,323906484629032960,323933624816787456,323913746231750656,323906398423490561,323946504752603136,323921167381577728,323928504523694080,323906482489921537,323916814490939393,323922279182520320,323915624835342337,323950257832603648,323924665074581504,323913403460636673,323925432548335616,323917413789868034,323944113764438016,323932515599515649,323924496144797697,323930212121317377,323918009901150208,323929272685318147,323911097818173440,323927822072684544,323924155391164416,323913403951357953,323919716039155712,323906484629032960,323933624816787456,323913746231750656,323906398423490561,323946504752603136,323921167381577728,323928504523694080,

6 := 324210976629067776,324038827578626048,324067437886713856,325063081929629696,324210976629067776,324038827578626048,324067437886713856,325063081929629696,

7 := 324117952695771137,324117950774775809,324117952695771137,324117950774775809,

8 := 324534627039588355,324070589214117888,324227779530985472,324002623571255296,324160230760005632,324188481918230528,324138055772561408,324460055724433408,324154782782742529,324179340675915779,324132606608306176,324125626284007424,324534627039588355,324070589214117888,324227779530985472,324002623571255296,324160230760005632,324188481918230528,324138055772561408,324460055724433408,324154782782742529,324179340675915779,324132606608306176,324125626284007424,

9 := 324219503401644033,324375472681148416,324366333829586945,324419757661097984,324358798179459073,324376801382133761,324219503401644033,324375472681148416,324366333829586945,324419757661097984,324358798179459073,324376801382133761,

10 := 324320247018573824,324327410344542208,324427314060484608,324428255794978816,324423829088784384,324327405743382529,324425542629732352,324427930631544832,324427596253253633,324320247018573824,324327410344542208,324427314060484608,324428255794978816,324423829088784384,324327405743382529,324425542629732352,324427930631544832,324427596253253633,

11 := 324348252763873283,324337983412847463,324382984012350293,324346553835868161,324310958418232934,324368024561909760,324373063433847463,324326985423982395,324334932112843721,324351038741901865,324326736198238763,324354128398432734,324348252763873283,324337983412847463,324382984012350293,324346553835868161,324310958418232934,324368024561909760,324373063433847463,324326985423982395,324334932112843721,324351038741901865,324326736198238763,324354128398432734,

12 := 324458862537228289,324380629779247106,324372750330363904,324418112688623616,324470253339959298,324458862537228289,324380629779247106,324372750330363904,324418112688623616,324470253339959298,

13 := 324432383115935744,324441544377634816,324446653635522560,324409861041438720,324439832715747328,324440178842296321,324441529030701057,324408861563944960,324445292130877440,324408472441585664,324439166496690176,324445289912074240,324452116863258624,324438127102013440,324409871061618689,324437436077854721,324432383115935744,324441544377634816,324446653635522560,324409861041438720,324439832715747328,324440178842296321,324441529030701057,324408861563944960,324445292130877440,324408472441585664,324439166496690176,324445289912074240,324452116863258624,324438127102013440,324409871061618689,324437436077854721,

14 := 324422817565257728,324423146381901825,324423506957852672,324423494098120705,324423160525103105,324423157106749441,324422817565257728,324423146381901825,324423506957852672,324423494098120705,324423160525103105,324423157106749441,

15 := 324451075501457408,324452781371060224,324448711860158464,324448359932887040,324448013999304704,324451075501457408,324452781371060224,324448711860158464,324448359932887040,324448013999304704,

16 := 324785129241341952,324790912431583232,324787508997144576,324785131569168384,324785122073251840,324799030859280384,324785120085176320,324785129241341952,324790912431583232,324787508997144576,324785131569168384,324785122073251840,324799030859280384,324785120085176320,

17 := 325061078994284544,325059351209443329,325059996704452608,325063478140346370,325059948511899648,325060608246562816,325059374814994434,325059598488829953,325061078994284544,325059351209443329,325059996704452608,325063478140346370,325059948511899648,325060608246562816,325059374814994434,325059598488829953,

18 := 325060145082163200,325060826346184707,325060138891350016,325060324992643072,325060154087309312,325060145082163200,325060826346184707,325060138891350016,325060324992643072,325060154087309312,

19 := 325166486215856129,325165164561330176,325175391654408192,325163043233337344,325171606253080577,325171288526163968,325170934115864576,325176378595102720,325178436660375552,325172960413503489,325170591164403712,325173634563964929,325162944931438592,325166486215856129,325165164561330176,325175391654408192,325163043233337344,325171606253080577,325171288526163968,325170934115864576,325176378595102720,325178436660375552,325172960413503489,325170591164403712,325173634563964929,325162944931438592,

20 := 325253670746849280,325253670780416000,325253327048822784,325253327640223744,325254353877352448,325253669408874496,325253670746849280,325253670780416000,325253327048822784,325253327640223744,325254353877352448,325253669408874496,

21 := 325329802032721920,325586012854185985,325338650076405761,325337627857387521,325342218942554112,325337623910559745,325335253487063041,325350010961141760,325337640322871296,325329802032721920,325586012854185985,325338650076405761,325337627857387521,325342218942554112,325337623910559745,325335253487063041,325350010961141760,325337640322871296,

22 := 325409910642835456,325410327183364096,325409910714138625,325410327296614400,325409910886117377,325409927193567232,325409992121389058,325409910588317696,325410327036571649,325409910827401217,325409996290527232,325409996726743041,325410327200145409,325409996529602560,325409996764479488,325409996965810176,325409910869327872,325410254697414658,325409910642835456,325410327183364096,325409910714138625,325410327296614400,325409910886117377,325409927193567232,325409992121389058,325409910588317696,325410327036571649,325409910827401217,325409996290527232,325409996726743041,325410327200145409,325409996529602560,325409996764479488,325409996965810176,325409910869327872,325410254697414658,

23 := 325702031912349696,325702375430057984,325702115865538560,325701932083716096,325701941478973440,325702271591665664,325702377279729664,325702956185948160,326228987233841152,325702632603803648,325702633174204416,325702285621612545,325702890297643008,325702626169729024,325702955896553472,325702958786441216,325702936036524032,325702711729348608,325702273684627456,325702269288992768,325702294287048705,325702287978795009,325702635405574145,325702363174285312,325701934273134594,325702294928764929,325702950859198464,325702031912349696,325702375430057984,325702115865538560,325701932083716096,325701941478973440,325702271591665664,325702377279729664,325702956185948160,326228987233841152,325702632603803648,325702633174204416,325702285621612545,325702890297643008,325702626169729024,325702955896553472,325702958786441216,325702936036524032,325702711729348608,325702273684627456,325702269288992768,325702294287048705,325702287978795009,325702635405574145,325702363174285312,325701934273134594,325702294928764929,325702950859198464,

24 := 325947343281205248,325927476566032385,325946633986641920,325947086145212416,325946283368013824,325946918842818560,325917940564250624,325947340777222145,325946659739668480,325947301745029121,325947343281205248,325927476566032385,325946633986641920,325947086145212416,325946283368013824,325946918842818560,325917940564250624,325947340777222145,325946659739668480,325947301745029121,