Clustering Tweets

EACH CLUSTER ID => represents the center for the datapoints / datasets below them.

SSD => a measurement of efficiency

K => Number of Clusters you want

Points => The IDs of each point the cluster, Of which one whill be chosen as CENTER

(CLUSTER ID)

Text: => The text represents the center of each cluster ( Most hit tweet words in the data, which represent the tweets in all of the cluster below)

**Working** :

All tweets in JSON format are parsed, and then taken into lists.

Then of the tweet data, stock words are removed and important data is extracted.

The above process is repeated for all the tweets.

Initially all the data is placed under one group( ONE CLUSTER)

Then for the number of clusters you want , data is gradually classified into the number of groups you wanted.

Then Jaccard distances (for all TWEETS) are used as a metric in separating the data( HOW CLOSE IS ONE SENTENCE RELATED TO ANOTHER), and then classified into the groups.

Sum of Squares of Distances between each and every tweet is identified , and the one which gives minimum value, is chosen as the CENTER of cluster. This is supposed to be minimum ( for each cluster) in order to efficiently classify the data.

The Jaccard distances and centers are fed to KMEANS, which would do the classification .

Due to some modifications in code, experimenting , I had to hard code the input URL, so it doesn’t take in from data file.

Change the hard coded URL in the CODE to your own URL.

**Execute in this way:**

**python tweets.py <numberOfClusters> <initialSeeds-file> <TweetsData-file> <output-flie>**

( Please url: <tweetsData file not used, instead the url of tweets is used, CODED)

ssd: 440.493475405 k = 2

K = 2

Iterations : 0 1

Cluster ID: 325409996764479488

text : set([u'RT', u'BREAKING', u'NEWS:', u'Boston', u'arrested.', u'@BostonGlobe:', u'suspect', u'terror', u'Marathon', u'bombing'])

Points : 323906397609791488 , 323906397735641088 , 323906397853073410 , 323906397618196483 , 324320247018573824 , 324787508997144576 , 324372750330363904 , 323906398352195585 , 324380629779247106 , 323906398230544385 , 325337627857387521 , 323929272685318147 , 324423146381901825 , 324428255794978816 , 325253670780416000 , 325409996764479488 , 325409927193567232 , 324368024561909760 , 323924496144797697 , 323932515599515649 , 324351038741901865 , 324425542629732352 , 323918009901150208 , 325175391654408192 , 324358798179459073 , 323906653651079168 , 323928504523694080 , 323913403951357953 , 324439832715747328 , 323910330315075584 , 323906398423490561 , 325409996965810176 , 323906650987692034 , 323906398314438656 , 324452116863258624 , 324423157106749441 , 323906656318676993 , 325409992121389058 , 325342218942554112 , 324229792834674689 , 325409910588317696 , 323906399149109248 , 323910330457669633 , 324348252763873283 , 324310958418232934 , 323906398993932289 , 324799030859280384 , 323906399295926273 , 325946918842818560 , 324785131569168384 , 325254353877352448 , 324451075501457408 , 323916814490939393 , 323955716392112128 , 325171288526163968 , 325337640322871296 , 323922279182520320 , 323906398176030720 , 325409996529602560 , 325947343281205248 , 325927476566032385 , 324423160525103105 , 323909308188344320 , 325059948511899648 , 325947301745029121 , 324375472681148416 , 324452781371060224 , 325060154087309312 , 323924665074581504 , 325335253487063041 , 324409871061618689 , 324438127102013440 , 324373063433847463 , 323907087551836160 , 325060324992643072 , 324132606608306176 , 324160230760005632 , 323906485249789952 , 324419757661097984 , 324334932112843721 , 324448359932887040 , 323906397962121216 , 323908795254312962 , 324408861563944960 , 323907771256938496 , 325171606253080577 , 323946504752603136 , 324219503401644033 , 324785120085176320 , 323915000567697409 , 324441529030701057 , 325063478140346370 , 323924155391164416 , 323933624816787456 , 323906398012461057 , 324422817565257728 , 325061078994284544 , 325409910827401217 , 323906483584655360 , 325060826346184707 , 324470253339959298 , 325163043233337344 , 325409910714138625 , 324210976629067776 , 325060138891350016 , 325946633986641920 , 323911097818173440 , 324427930631544832 , 324427596253253633 , 325410327296614400 , 324790912431583232 , 325253327640223744 , 325350010961141760 , 323950257832603648 , 325410327200145409 , 325947086145212416 , 325946283368013824 , 324326736198238763 , 324423829088784384 , 324445289912074240 , 324423506957852672 , 325173634563964929 , 325170934115864576 , 323906657333682176 , 323919716039155712 , 324423494098120705 , 325409910869327872 , 324408472441585664 , 323906482489921537 , 323925264352546816 , 325166486215856129 , 325178436660375552 , 324427314060484608 , 323913746231750656 , 323932094190439874 , 323906484629032960 , 325170591164403712 , 323907258301939713 , 324366333829586945 , 323916051614138368 , 324326985423982395 , 325917940564250624 , 324409861041438720 , 323917413789868034 , 325059351209443329 , 323923559799996417 , 325409996726743041 , 323927822072684544 , 325165164561330176 , 325409910886117377 , 325253670746849280 , 324382984012350293 , 323915624835342337 , 324445292130877440 , 325059996704452608 , 323920146454425600 , 325253327048822784 , 325410327183364096 , 324337983412847463 , 325060608246562816 , 323913403460636673 , 324038827578626048 , 325338650076405761 , 324327410344542208 , 325172960413503489 , 323921167381577728 , 325586012854185985 , 323911610236293120 , 324785129241341952 , 325337623910559745 , 325176378595102720 , 325059598488829953 , 325409996290527232 , 325253669408874496 , 324376801382133761 , 325946659739668480 , 325329802032721920 , 324448711860158464 , 325059374814994434 , 323930212121317377 , 323906398826164225 , 325162944931438592 , 325063081929629696 , 323908795396943872 , 324446653635522560 , 323944113764438016 , 324067437886713856 , 324327405743382529 , 325060145082163200 , 323906651209994241 , 323906567294562306 , 324448013999304704 , 325410327036571649 , 323925432548335616 , 324437436077854721 , 323963901769297921 , 324440178842296321 , 324418112688623616 , 325409910642835456 , 323906399300100096 , 324441544377634816 , 323908455545049088 , 325410254697414658 , 323921510282702848 , 324785122073251840 , 324354128398432734 , 324458862537228289 , 325947340777222145 , 324226045052071936 , 324432383115935744 , 324346553835868161 , 324439166496690176 ,

Cluster ID: 325702285621612545

text : set([u'Aerial', u'Show', u'Boston', u'http://t.co/9AaPTLXQQE', u'Suspect', u'In', u'Images', u'New', u'Boat', u'Bombing'])

Points : 325702271591665664 , 325702031912349696 , 325702626169729024 , 325702955896553472 , 325702890297643008 , 324070589214117888 , 325702363174285312 , 325702633174204416 , 324125626284007424 , 325702936036524032 , 324227779530985472 , 325702711729348608 , 324117952695771137 , 325702956185948160 , 324460055724433408 , 325702273684627456 , 325702269288992768 , 325702377279729664 , 324154782782742529 , 325702375430057984 , 324117950774775809 , 325701941478973440 , 324179340675915779 , 324188481918230528 , 325702285621612545 , 324138055772561408 , 325702950859198464 , 325702635405574145 , 324002623571255296 , 325701932083716096 , 325702115865538560 , 325702287978795009 , 325702294928764929 , 325701934273134594 , 326228987233841152 , 324534627039588355 , 325702632603803648 , 325702958786441216 , 325702294287048705 ,

ssd: 622.45009307 k = 3

K = 3

Iterations : 0 1 2

Cluster ID: 325409996764479488

text : set([u'RT', u'BREAKING', u'NEWS:', u'Boston', u'arrested.', u'@BostonGlobe:', u'suspect', u'terror', u'Marathon', u'bombing'])

Points : 324320247018573824 , 324787508997144576 , 324372750330363904 , 324380629779247106 , 325337627857387521 , 323929272685318147 , 324423146381901825 , 324428255794978816 , 325253670780416000 , 325409996764479488 , 325409927193567232 , 323924496144797697 , 323932515599515649 , 324425542629732352 , 323918009901150208 , 325175391654408192 , 324358798179459073 , 323906653651079168 , 323928504523694080 , 323913403951357953 , 324439832715747328 , 323910330315075584 , 323906398423490561 , 325409996965810176 , 323906650987692034 , 324452116863258624 , 324423157106749441 , 325409992121389058 , 325342218942554112 , 324229792834674689 , 325409910588317696 , 323910330457669633 , 324799030859280384 , 325946918842818560 , 324785131569168384 , 325254353877352448 , 324451075501457408 , 323916814490939393 , 323955716392112128 , 325171288526163968 , 325337640322871296 , 323922279182520320 , 323906398176030720 , 325409996529602560 , 325947343281205248 , 325927476566032385 , 324423160525103105 , 323909308188344320 , 325059948511899648 , 325947301745029121 , 324375472681148416 , 324452781371060224 , 325060154087309312 , 323924665074581504 , 325335253487063041 , 324409871061618689 , 324438127102013440 , 325060324992643072 , 324132606608306176 , 324160230760005632 , 324419757661097984 , 324448359932887040 , 324408861563944960 , 325171606253080577 , 323946504752603136 , 324219503401644033 , 324785120085176320 , 324441529030701057 , 325063478140346370 , 323924155391164416 , 323933624816787456 , 324422817565257728 , 325061078994284544 , 325409910827401217 , 325060826346184707 , 324470253339959298 , 325163043233337344 , 325409910714138625 , 324210976629067776 , 325060138891350016 , 325946633986641920 , 323911097818173440 , 324427930631544832 , 324427596253253633 , 325410327296614400 , 324790912431583232 , 325253327640223744 , 325350010961141760 , 323950257832603648 , 325410327200145409 , 325947086145212416 , 325946283368013824 , 324423829088784384 , 324445289912074240 , 324423506957852672 , 325173634563964929 , 325170934115864576 , 323906657333682176 , 323919716039155712 , 324423494098120705 , 325409910869327872 , 324408472441585664 , 323906482489921537 , 325166486215856129 , 325178436660375552 , 324427314060484608 , 323913746231750656 , 323932094190439874 , 323906484629032960 , 325170591164403712 , 323907258301939713 , 324366333829586945 , 325917940564250624 , 324409861041438720 , 323917413789868034 , 325059351209443329 , 325409996726743041 , 323927822072684544 , 325165164561330176 , 325409910886117377 , 325253670746849280 , 323915624835342337 , 324445292130877440 , 325059996704452608 , 325253327048822784 , 325410327183364096 , 325060608246562816 , 323913403460636673 , 324038827578626048 , 325338650076405761 , 324327410344542208 , 325172960413503489 , 323921167381577728 , 325586012854185985 , 324785129241341952 , 325337623910559745 , 325176378595102720 , 325059598488829953 , 325409996290527232 , 325253669408874496 , 324376801382133761 , 325946659739668480 , 325329802032721920 , 324448711860158464 , 325059374814994434 , 323930212121317377 , 325162944931438592 , 325063081929629696 , 324446653635522560 , 323944113764438016 , 324067437886713856 , 324327405743382529 , 325060145082163200 , 323906651209994241 , 323906567294562306 , 324448013999304704 , 325410327036571649 , 323925432548335616 , 324437436077854721 , 323963901769297921 , 324440178842296321 , 324418112688623616 , 325409910642835456 , 324441544377634816 , 325410254697414658 , 324785122073251840 , 324458862537228289 , 325947340777222145 , 324226045052071936 , 324432383115935744 , 324439166496690176 ,

Cluster ID: 323906397735641088

text : set([u'RT', u'finish', u'run', u'@NBCSN:', u'Mass', u'Hospital', u'Reports', u'General', u'#PrayforBoston', u'crossed', u'Marathon', u'victims', u'line', u'continued', u'Runners', u'blood'])

Points : 323906397609791488 , 323906397735641088 , 323906397853073410 , 323906397618196483 , 323906398352195585 , 323906398230544385 , 324368024561909760 , 324351038741901865 , 323906398314438656 , 323906656318676993 , 323906399149109248 , 324348252763873283 , 324310958418232934 , 323906398993932289 , 323906399295926273 , 324373063433847463 , 323907087551836160 , 323906485249789952 , 324334932112843721 , 323906397962121216 , 323908795254312962 , 323907771256938496 , 323915000567697409 , 323906398012461057 , 323906483584655360 , 324326736198238763 , 323925264352546816 , 323916051614138368 , 324326985423982395 , 323923559799996417 , 324382984012350293 , 323920146454425600 , 324337983412847463 , 323911610236293120 , 323906398826164225 , 323908795396943872 , 323906399300100096 , 323908455545049088 , 323921510282702848 , 324354128398432734 , 324346553835868161 ,

Cluster ID: 325702285621612545

text : set([u'Aerial', u'Show', u'Boston', u'http://t.co/9AaPTLXQQE', u'Suspect', u'In', u'Images', u'New', u'Boat', u'Bombing'])

Points : 325702271591665664 , 325702031912349696 , 325702626169729024 , 325702955896553472 , 325702890297643008 , 324070589214117888 , 325702363174285312 , 325702633174204416 , 324125626284007424 , 325702936036524032 , 324227779530985472 , 325702711729348608 , 324117952695771137 , 325702956185948160 , 324460055724433408 , 325702273684627456 , 325702269288992768 , 325702377279729664 , 324154782782742529 , 325702375430057984 , 324117950774775809 , 325701941478973440 , 324179340675915779 , 324188481918230528 , 325702285621612545 , 324138055772561408 , 325702950859198464 , 325702635405574145 , 324002623571255296 , 325701932083716096 , 325702115865538560 , 325702287978795009 , 325702294928764929 , 325701934273134594 , 326228987233841152 , 324534627039588355 , 325702632603803648 , 325702958786441216 , 325702294287048705 ,

ssd: 839.817575218 k = 4

K = 4

Iterations : 0 1 2

Cluster ID: 324368024561909760

text : set([u'RT', u'#BostonMarathon', u'victims', u'For', u'@SiRHaPo:', u'receive', u'@\_BostonMarathon', u'#PrayForBoston', u'donate', u'retweet', u'$1'])

Points : 324372750330363904 , 324380629779247106 , 324368024561909760 , 324351038741901865 , 324358798179459073 , 324348252763873283 , 324310958418232934 , 325059948511899648 , 324373063433847463 , 323906485249789952 , 324419757661097984 , 324334932112843721 , 324219503401644033 , 323915000567697409 , 325063478140346370 , 325061078994284544 , 323906483584655360 , 324470253339959298 , 324210976629067776 , 324326736198238763 , 323925264352546816 , 324366333829586945 , 323916051614138368 , 324326985423982395 , 325059351209443329 , 323923559799996417 , 324382984012350293 , 325059996704452608 , 323920146454425600 , 324337983412847463 , 325060608246562816 , 324038827578626048 , 323911610236293120 , 325059598488829953 , 325059374814994434 , 325063081929629696 , 324067437886713856 , 324418112688623616 , 323921510282702848 , 324354128398432734 , 324458862537228289 , 324346553835868161 ,

Cluster ID: 325253670780416000

text : set([u'RT', u'Do', u'approach.', u'Police', u'seeking', u'#WANTED:', u'\u201999', u'Honda', u'-', u'Possible', u'http://t.co/11eRTJdtaZ', u'Plate:', u'Color', u'car.', u'suspect', u'316-ES9,', u'@Boston\_Police:', u'MA', u'Gray.', u'CRV,'])

Points : 325253670780416000 , 325254353877352448 , 325253327640223744 , 325917940564250624 , 325253670746849280 , 325253327048822784 , 325253669408874496 ,

Cluster ID: 325409996764479488

text : set([u'RT', u'BREAKING', u'NEWS:', u'Boston', u'arrested.', u'@BostonGlobe:', u'suspect', u'terror', u'Marathon', u'bombing'])

Points : 323906397609791488 , 323906397735641088 , 323906397853073410 , 323906397618196483 , 324320247018573824 , 324787508997144576 , 323906398352195585 , 323906398230544385 , 325337627857387521 , 323929272685318147 , 324423146381901825 , 324428255794978816 , 325409996764479488 , 325409927193567232 , 323924496144797697 , 323932515599515649 , 324425542629732352 , 323918009901150208 , 325175391654408192 , 323906653651079168 , 323928504523694080 , 323913403951357953 , 324439832715747328 , 323910330315075584 , 323906398423490561 , 325409996965810176 , 323906650987692034 , 323906398314438656 , 324452116863258624 , 324423157106749441 , 323906656318676993 , 325409992121389058 , 325342218942554112 , 324229792834674689 , 325409910588317696 , 323906399149109248 , 323910330457669633 , 323906398993932289 , 324799030859280384 , 323906399295926273 , 325946918842818560 , 324785131569168384 , 324451075501457408 , 323916814490939393 , 323955716392112128 , 325171288526163968 , 325337640322871296 , 323922279182520320 , 323906398176030720 , 325409996529602560 , 325947343281205248 , 325927476566032385 , 324423160525103105 , 323909308188344320 , 325947301745029121 , 324375472681148416 , 324452781371060224 , 325060154087309312 , 323924665074581504 , 325335253487063041 , 324409871061618689 , 324438127102013440 , 323907087551836160 , 325060324992643072 , 324132606608306176 , 324160230760005632 , 324448359932887040 , 323906397962121216 , 323908795254312962 , 324408861563944960 , 323907771256938496 , 325171606253080577 , 323946504752603136 , 324785120085176320 , 324441529030701057 , 323924155391164416 , 323933624816787456 , 323906398012461057 , 324422817565257728 , 325409910827401217 , 325060826346184707 , 325163043233337344 , 325409910714138625 , 325060138891350016 , 325946633986641920 , 323911097818173440 , 324427930631544832 , 324427596253253633 , 325410327296614400 , 324790912431583232 , 325350010961141760 , 323950257832603648 , 325410327200145409 , 325947086145212416 , 325946283368013824 , 324423829088784384 , 324445289912074240 , 324423506957852672 , 325173634563964929 , 325170934115864576 , 323906657333682176 , 323919716039155712 , 324423494098120705 , 325409910869327872 , 324408472441585664 , 323906482489921537 , 325166486215856129 , 325178436660375552 , 324427314060484608 , 323913746231750656 , 323932094190439874 , 323906484629032960 , 325170591164403712 , 323907258301939713 , 324409861041438720 , 323917413789868034 , 325409996726743041 , 323927822072684544 , 325165164561330176 , 325409910886117377 , 323915624835342337 , 324445292130877440 , 325410327183364096 , 323913403460636673 , 325338650076405761 , 324327410344542208 , 325172960413503489 , 323921167381577728 , 325586012854185985 , 324785129241341952 , 325337623910559745 , 325176378595102720 , 325409996290527232 , 324376801382133761 , 325946659739668480 , 325329802032721920 , 324448711860158464 , 323930212121317377 , 323906398826164225 , 325162944931438592 , 323908795396943872 , 324446653635522560 , 323944113764438016 , 324327405743382529 , 325060145082163200 , 323906651209994241 , 323906567294562306 , 324448013999304704 , 325410327036571649 , 323925432548335616 , 324437436077854721 , 323963901769297921 , 324440178842296321 , 325409910642835456 , 323906399300100096 , 324441544377634816 , 323908455545049088 , 325410254697414658 , 324785122073251840 , 325947340777222145 , 324226045052071936 , 324432383115935744 , 324439166496690176 ,

Cluster ID: 325702285621612545

text : set([u'Aerial', u'Show', u'Boston', u'http://t.co/9AaPTLXQQE', u'Suspect', u'In', u'Images', u'New', u'Boat', u'Bombing'])

Points : 325702271591665664 , 325702031912349696 , 325702626169729024 , 325702955896553472 , 325702890297643008 , 324070589214117888 , 325702363174285312 , 325702633174204416 , 324125626284007424 , 325702936036524032 , 324227779530985472 , 325702711729348608 , 324117952695771137 , 325702956185948160 , 324460055724433408 , 325702273684627456 , 325702269288992768 , 325702377279729664 , 324154782782742529 , 325702375430057984 , 324117950774775809 , 325701941478973440 , 324179340675915779 , 324188481918230528 , 325702285621612545 , 324138055772561408 , 325702950859198464 , 325702635405574145 , 324002623571255296 , 325701932083716096 , 325702115865538560 , 325702287978795009 , 325702294928764929 , 325701934273134594 , 326228987233841152 , 324534627039588355 , 325702632603803648 , 325702958786441216 , 325702294287048705 ,

ssd: 5527.72198503 k = 25