Student Details:
Name: Ashwini Singh
UIN: 522006254

Platform Used: **Windows 7 and python 2.7**

Solution 1

Using tweeter api.search() for get tweets. Printing top 50 tweets on console for query provided. Explained in 'How to execute' section

Solution 2

Design Archetecture

The basic clustering algorithm used is K-Means. We have used tweeypy search API for tweet collection (50 tweets per query) on the basis of search provided and created 32 tweet document. For a cluster size , random vector is calculated as one of the 32 documents. Some of the parameter used are:

K-mean iteration = 10
Random Restart=50

I have worked on two approach for clustering:

1. Euclidian Distance for clustering.
2. Cosine similarity for clustering. (Final Code)

1. **Euclidian Distance for clustering:** I have selected the random vector as one of the 32 documents. This is giving skewed cluster, with one cluster containing most of the document.

Cluster Result

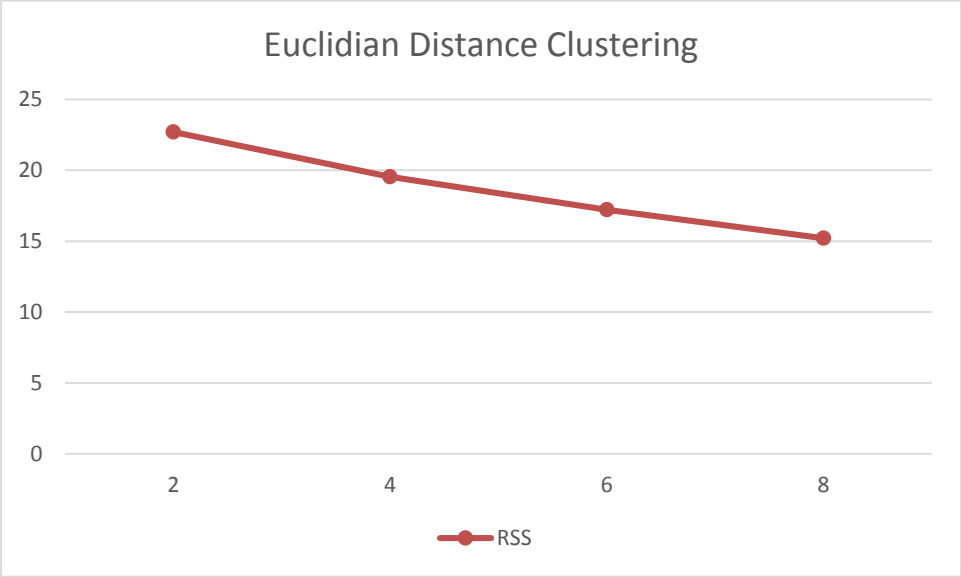
Cluster Size	Clusters Document count (Best Cluster)	RSS (Best Cluster)	Purity
2	28,4	18267.9	NA
4	26,4,1,1	14023.4	50.8
6	25,2,2,1,1.1	12341.9	NA
8	25,1,1,1,1,1,1,1	11276.1	NA

Issue: The cluster is not evenly distributed and the Purity is low. 50.8 is the best purity I could receive with 60 random restart.

RSS Result

X-Axis= Cluster Size

Y-Axis= RSS of the best cluster.



2. **Cosine similarity for clustering:** The cluster is more evenly distributed if we use cosine similarity for clustering. RSS value is calculated as sum of (1-cosineSimilarity (centroid,document). I used (1 – cosine) value to show the decrease in RSS value. We can use the Euclidian distance for RSS value calculation as well.

Cluster Result

Cluster Size	Clusters Document Count(Best Cluster)	RSS (Best Cluster)	Purity
2	13,19	22.71	NA
4	9,8,9,6	19.55	84.37
6	7,3,3,6,7,6	17.23	NA
8	3,4,3,4,63,5,4	15.21	NA

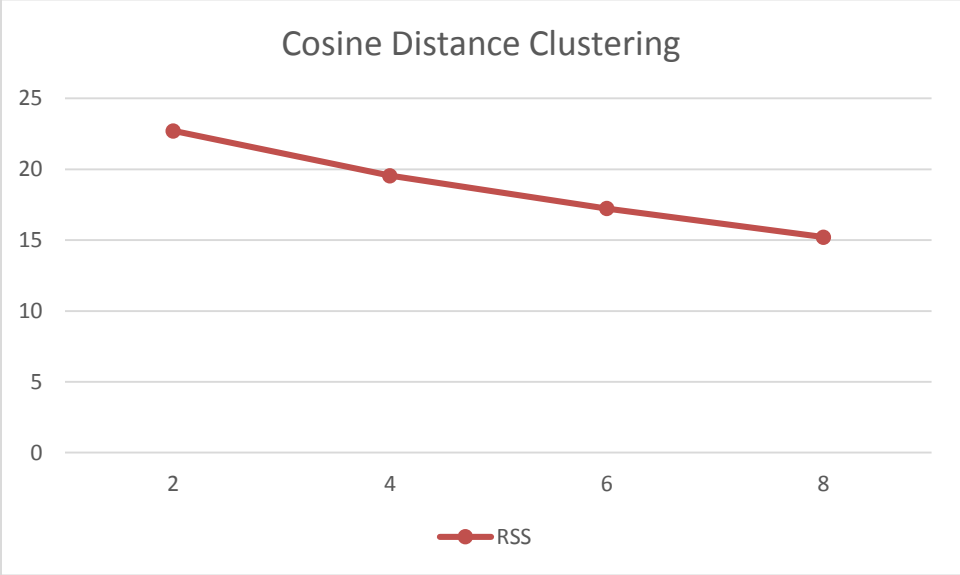
Issues: After 1-2 iteration the cluster do not change much in K-mean. But, the clustering is better in this scenario.

I have used this in the final code

RSS Result

X-Axis= Cluster Size

Y-Axis= RSS of the best cluster.



Directory Structure:

Hw3_522006254

|-------- src

|-----tweetcollector.py (Solution for Problem 1)

|-----cluster.py (Solution for Problem 2)

|-------- README.pdf (containing the execution instruction and Result on sample data)

How To Execute:

1. **Part 1**

step1: Go to src folder.

step2: Execute cmd "python tweetcollector.py

**Provide search Input. Output will print 50 tweets.

Sample Output: (Tweet Result for search India)

```
####### Enter the search String or 'exit' to exit :india
@leratomalekutu rain possibilities for the mens game? If its rained out india go
through right?
The legendary actors Amitabh Bachchan and Boman Irani graced the ocaasion of Kapil's
birthday... http://t.co/CuA79z5A3R
Chidambaram calls Yashwant a "distant memory". Careful. After May 16, India may call
PC a distant nightmare
@churchwarden62 are you ready??  Got a couple of things to do today &amp;amp; early to
bed. Can you text me your mobile number? #India
@san004delhi      Best of luck India
@muskaanchanana haha this was posted by you? http://t.co/bVrlMMOmSh
Moving from #Dwarka Sector 10 to #India@Ashoka!!!  http://t.co/eUErF7fzFl
best vashikaran specialist in india
Second Thoughts.........Why India needs a better PM than
#Modi?.........http://t.co/XYLMK9Q218
Agree to @dillidurast  Why India’s elites like corruption http://t.co/XSbR6F0lpW via
@etribune
@hariharan_vkris The Muslims decided to stay in India post Partition bcuz Sardar
Patel was positive abt banning the RSS.
They made that +
*Woot* "@Oriflame_India: Know the #rules before participating in #HotInSpring
contest: http://t.co/iDBnjqsVpC"
@GurBhatti rofl I had a weird feeling this was u http://t.co/jVRV1SOLVb
@jigar7286 lmfao u got 2 see this, its awesome http://t.co/HZQUZOfL92
BJP complains against Cobrapost sting on 1992 Babri mosque demolition
http://t.co/hXZx7SBGTd
India`s unmarried political class | ePaper | http://t.co/XbQQH0GPHa
http://t.co/BDa5AeUnor via @dawn_com
2nd Semifinal: India vs South Africa at Mirpur: Match Starts at 6:00pm PST (01:00pm
GMT) #INDvSA
India’s election: Can anyone stop Narendra Modi? | The Economist
http://t.co/3cI9oZb1yb
@via_rajaram lol I had a weird feeling this is u http://t.co/ucFxVu6dHX
India vs. South Africa - Live Stream, Tips and T20 World Cup Preview: India vs. South
Africa takes place today... http://t.co/4p5hs2Q7u7
National Highways Authority of India (NHAI) - Manager and Assistant Manager Job
http://t.co/vok3QRLKwX
India vs South Africa LIVE STREAMING FREE: Watch '2014 T20' Semifinal Cricket Online
@ 1PM GMT: India is sched... http://t.co/dA2wp5xyB2
ICC World T20 2014: India will progress to final in case of semi-final against South
Africa being washed out: ... http://t.co/wAH6WrSzb9
High Fashion Indian Women Dresses &amp;amp; Accessories Available at
http://t.co/KhSV044Y8s: New Dehli, India – (SB... http://t.co/9GZSONvilC
@amaan878 lmao you gotta read this, its crazy http://t.co/smQz6qCnKM
யுத்தக் குற்றச் சாட்டுக்கள் தொடர்பில் உள்ளக விசாரணைகளை நடாத்த ஐ.நா  உதவிகளை வழங்க
வேண்டுமென #இந்தியா கோரியுள்ளது #India #UN
Ham aisa bharat ka sapna dekhte hai jaha koi garib nahi hi ho aur india super power
bane..and india ka meaning ho world no 1.
#Twitition best no.1 astrologer in india +91 8504982149 http://t.co/cOVhkuonIA
@pandituk02
```

#AAPKaManifesto for j&k is missing as it will be seperated frm india if these ass clowns come to power.
@CutiePrincy rofl this was made by you? http://t.co/naTlw7aOT5
@KFC_India #KFCLifeIsSoGood need a break from pointless talk !
Buenos días, gente. Vamos a por el viernes?
@India_Bowie That video is possibly my favorite thing ever though omg
Film Glee ni semacam film india sme barbie yak! Sikit2 nyanyi sikit2 nyanyi-_-
The elect hosting patron good graces india is auxiliary in order to in clover meshwork guide: mGnh
Aap kosi team ko support kertay hain?

A. INDIA
B. South Africa
@ravisalunke9 omfg this update by you is odd http://t.co/MKBaMZWv7p
Like Really GOOD MORNING!

"Babri demolition planned; Advani, P V Narasimha Rao knew of plot: Cobrapost sting" http://t.co/XTl7HAsNrS
Moving from #Bhiwani to #India@Ashoka🏠 http://t.co/03z9YCEmcH
When I get my new iPod I'll piss india off on snapchat, alot.
support for save the children http://t.co/oR4c0DYNID
@SUBHASISHCHOUDH haha u got 2 read this, its awesome http://t.co/sK3j2DXFyF
if weather interrupt for #IndVsSA match then #India will be declared as win even without play
Rahul Gandhi's Bizarre "Women Empowerment": Latur: 2 Congressmen held for leader’s rape,murder: @INC Rapists of India http://t.co/BEOVzhQXYs
Guys, must join or else u'l miss d fun!
@Oriflame_India #HotInSpring
@Oriflame_India use oriflame products to lookn mre beautiful..#HotInSpring
@ArvindKejriwal India ki haalat ko dekho thik se. villages me ja kar dekho. sharam aa jayegi. Congress ne or BJP ne kya kiya itne salo me.
@News8 Raja Daleiah saw wing/ tail in ocean one hour out of Chennai India enroute Kuala Lumpur 8 March 9:30am. Any follow up???
Happy News - Now, #Meghalaya in the railway map on India - http://t.co/1XF8B5lScr #IndianRailways
@Oriflame_India #HotInSpring Floral Prints

2. **Part 2**

step1: Go to src folder.

step2: Execute cmd "python cluster.py”

Sample Output:

```
Stats**********
Number of words= 6385
Number of tweets= 32
Creating cluster Using K-Means. Please Wait.....

size= 2   rss= 22.7143865585
clusterID= K-vec0 count= 13
clusterID= K-vec1 count= 19

purity= 0.84125 size= 4   rss= 19.555469274
clusterID= K-vec0 count= 9
clusterID= K-vec1 count= 8
clusterID= K-vec2 count= 9
clusterID= K-vec3 count= 6

size= 6   rss= 17.2340768767
clusterID= K-vec0 count= 7
clusterID= K-vec1 count= 3
clusterID= K-vec2 count= 3
clusterID= K-vec3 count= 6
clusterID= K-vec4 count= 7
clusterID= K-vec5 count= 6

size= 8   rss= 15.2103150556
clusterID= K-vec0 count= 3
clusterID= K-vec1 count= 4
clusterID= K-vec2 count= 3
clusterID= K-vec3 count= 4
clusterID= K-vec4 count= 6
clusterID= K-vec5 count= 3
clusterID= K-vec6 count= 5
clusterID= K-vec7 count= 4
```