DATASCI W261, Machine Learning at Scale

Assignement: Midterm Exam

Lei Yang (mailto:leiyang@berkeley.edu)

Due: 2016-03-02, 6PM PST

In [ ]:

In [15]:
```
%%writefile kltext.txt
1.Data Science is an interdisciplinary field about processes and systems to extract knowledge or i
nsights from large volumes of data in various forms (data in various forms, data in various forms,
data in various forms), either structured or unstructured,[1][2] which is a continuation of some o
f the data analysis fields such as statistics, data mining and predictive analytics, as well as Kn
owledge Discovery in Databases.
2.Machine learning is a subfield of computer science[1] that evolved from the study of pattern rec
ognition and computational learning theory in artificial intelligence.[1] Machine learning explore
s the study and construction of algorithms that can learn from and make predictions on data.[2] Su
ch algorithms operate by building a model from example inputs in order to make data-driven predict
ions or decisions,[3]:2 rather than following strictly static program instructions.
```

```
Overwriting kltext.txt
```

In [16]:
```
import numpy as np
np.log(3)
```

Out[16]:
```
1.0986122886681098
```

Question 6, 7, 8

- define a dummy reducer_init, and stop execution after it, to see sorted mapper output, for question 7
- swap mapper1_smooth for mapper for question 8

In [19]:
```
%%writefile kldivergence.py
from mrjob.job import MRJob
from math import log
import re
import numpy as np
class kldivergence(MRJob):
    def mapper1(self, _, line):
        index = int(line.split('.',1)[0])
        letter_list = re.sub(r"[^A-Za-z]+", '', line).lower()
        count = {}
        for l in letter_list:
            if count.has_key(l):
                count[l] += 1
            else:
```

```
                count[l] = 1
        for key in count:
            yield key, [index, count[key]*1.0/len(letter_list)]

    def mapper1_smooth(self, _, line):
        index = int(line.split('.',1)[0])
        letter_list = re.sub(r"[^A-Za-z]+", '', line).lower()
        count = {}
        for l in letter_list:
            if count.has_key(l):
                count[l] += 1
            else:
                count[l] = 1
        for key in count:
            yield key, [index, (1+count[key]*1.0)/(24+len(letter_list))]

    # add this thing here as reducer_init, to just see the sorted mapper output
    def temp(self):
        b=0

    def reducer1(self, key, values):
        #Fill in your code
        # probability holder inline with line index, so that prob[1], prob[2] are P_1, P_2
        prob = [0, 0, 0]
        # use index to control probablity, as there is no guarantee they arrive as 1, 2 in order
        for v in values:
            index, p = v[0], v[1]
            prob[index] = p
        yield None, prob[1] * log(prob[1]/prob[2])

    # added, otherwise weird error it won't run
    def mapper2(self, key, value):
        yield key, value

    def reducer2(self, key, values):
        kl_sum = 0
        for value in values:
            kl_sum = kl_sum + value
        yield None, kl_sum

    def steps(self):
        return [self.mr(###### switch mapper for (non)smoothing ######
                        #mapper=self.mapper1
                        mapper=self.mapper1_smooth
                        ,reducer_init=self.temp
                        ,reducer=self.reducer1
                       )

                , self.mr(mapper=self.mapper2, reducer=self.reducer2)
               ]

if __name__ == '__main__':
    kldivergence.run()
```

```
Overwriting kldivergence.py
```

Execution for Question 6

In [18]:
```
!python kldivergence.py kltext.txt
```

```
using configs in /Users/leiyang/.mrjob.conf
creating tmp directory /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160
303.015349.615427
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.6
15427/step-0-mapper_part-00000
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
Counters from step 1:
  (no counters found)
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.6
15427/step-0-mapper-sorted
```

```
> sort /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.61542
7/step-0-mapper_part-00000
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.6
15427/step-0-reducer_part-00000
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
Counters from step 1:
  (no counters found)
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.6
15427/step-1-mapper_part-00000
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
Counters from step 2:
  (no counters found)
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.6
15427/step-1-mapper-sorted
> sort /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.61542
7/step-1-mapper_part-00000
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.6
15427/step-1-reducer_part-00000
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly instead.
Counters from step 2:
  (no counters found)
Moving /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160303.015349.61542
7/step-1-reducer_part-00000 -> /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiya
ng.20160303.015349.615427/output/part-00000
Streaming final output from /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.
20160303.015349.615427/output
null    0.08088278445318145
removing tmp directory /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/kldivergence.leiyang.20160
303.015349.615427
```

Execution for question 8

In [20]:
```
%load_ext autoreload
%autoreload 2
from kldivergence import kldivergence

mr_job = kldivergence(args=['kltext.txt'])
with mr_job.make_runner() as runner:
    runner.run()
    # stream_output: get access of the output
    for line in runner.stream_output():
        print mr_job.parse_output_line(line)
```

```
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
```

```
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
WARNING:mrjob.job:mr() is deprecated and will be removed in v0.6.0. Use mrjob.step.MRStep directly
instead.
(None, 0.06726997279170038)
```

MrJob class for Kmeans - Question 10 & 11

If you want to change the code, please edit Kmeans.py directly

In [4]:

```
%%writefile Kmeans.py
from numpy import argmin, array, random
from mrjob.job import MRJob
from mrjob.step import MRJobStep
from itertools import chain
from math import sqrt

#Calculate find the nearest centroid for data point
def MinDist(datapoint, centroid_points):
    datapoint = array(datapoint)
    centroid_points = array(centroid_points)
    diff = datapoint - centroid_points
    diffsq = diff**2

    distances = (diffsq.sum(axis = 1))**0.5
    # Get the nearest centroid for each instance
    min_idx = argmin(distances)
    return min_idx

#Check whether centroids converge
def stop_criterion(centroid_points_old, centroid_points_new,T):
    oldvalue = list(chain(*centroid_points_old))
    newvalue = list(chain(*centroid_points_new))
    Diff = [abs(x-y) for x, y in zip(oldvalue, newvalue)]
    Flag = True
    for i in Diff:
        if(i>T):
            Flag = False
            break
    return Flag


class MRKmeans(MRJob):
    centroid_points=[]
    k=3
    def steps(self):
        return [
            MRJobStep(mapper_init = self.mapper_init, mapper=self.mapper,combiner = self.combiner,
reducer=self.reducer)
               ]
    #load centroids info from file
    def mapper_init(self):
        self.centroid_points = [map(float,s.split('\n')[0].split(',')) for s in open("Centroids.tx
t").readlines()]
        open('Centroids.txt', 'w').close()
    #load data and output the nearest centroid index and data point
    def mapper(self, _, line):
        D = (map(float,line.split(',')))
        ######################## let's weight the input ########################
        norm = sqrt(sum([x**2 for x in D]))
        D = [x/norm for x in D]
        idx = MinDist(D,self.centroid_points)
        yield int(idx), (D[0],D[1],1)
    #Combine sum of data points locally
```

```
    def combiner(self, idx, inputdata):
        sumx = sumy = num = 0
        for x,y,n in inputdata:
            num = num + n
            sumx = sumx + x
            sumy = sumy + y
        yield int(idx),(sumx,sumy,num)
    #Aggregate sum for each cluster and then calculate the new centroids
    def reducer(self, idx, inputdata):
        centroids = []
        num = [0]*self.k
        distances = 0
        for i in range(self.k):
            centroids.append([0,0])
        for x, y, n in inputdata:
            num[idx] = num[idx] + n
            centroids[idx][0] = centroids[idx][0] + x
            centroids[idx][1] = centroids[idx][1] + y
        centroids[idx][0] = centroids[idx][0]/num[idx]
        centroids[idx][1] = centroids[idx][1]/num[idx]
        with open('Centroids.txt', 'a') as f:
            f.writelines(str(centroids[idx][0]) + ',' + str(centroids[idx][1]) + '\n')
        yield idx,(centroids[idx][0],centroids[idx][1])

if __name__ == '__main__':
    MRKmeans.run()
```

```
Overwriting Kmeans.py
```

Driver: Execution for Question 10

Generate random initial centroids

New Centroids = initial centroids

While(1):

- Cacluate new centroids
- stop if new centroids close to old centroids
- Updates centroids

In [5]:
```
from numpy import random, array
from Kmeans import MRKmeans, stop_criterion
mr_job = MRKmeans(args=['Kmeandata.csv', '--file', 'Centroids.txt'])

#Geneate initial centroids
centroid_points = [[0,0],[6,3],[3,6]]
k = 3
with open('Centroids.txt', 'w+') as f:
        f.writelines(','.join(str(j) for j in i) + '\n' for i in centroid_points)

# Update centroids iteratively
for i in range(10):
    # save previous centoids to check convergency
    centroid_points_old = centroid_points[:]
    print "iteration"+str(i+1)+":"
    with mr_job.make_runner() as runner:
        runner.run()
        # stream_output: get access of the output
        for line in runner.stream_output():
            key,value =  mr_job.parse_output_line(line)
            print key, value
            centroid_points[key] = value
    print "\n"
    i = i + 1
print "Centroids\n"
print centroid_points
```

```
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
```

```
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
iteration1:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-3.344726378997624, 0.3375985510805811]
1 [5.379067911319127, 0.1544680529517142]
2 [0.2428827627020568, 5.350519186138142]


iteration2:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.938524015701945, 0.043216587887174585]
1 [5.040232716088853, -0.02629422997828942]
2 [0.08609737928171676, 5.025145679728707]


iteration3:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration4:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration5:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration6:
```

```
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration7:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration8:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration9:
0
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
WARNING:mrjob.step:MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


iteration10:
0 [-4.985805688899424, 0.0009376094363627237]
1 [5.040232716088853, -0.02629422997828942]
2 [0.053065423788148436, 4.987793423944292]


Centroids

[[-4.985805688899424, 0.0009376094363627237], [5.040232716088853, -0.02629422997828942], [0.053065
423788148436, 4.987793423944292]]
```

In [ ]:

execution for question 11

In [13]:

```
%%writefile Quesion11.py
```

```
from numpy import argmin, array, random
from mrjob.job import MRJob
from mrjob.step import MRJobStep
from itertools import chain
from math import sqrt

def MinDist(datapoint, centroid_points):
    datapoint = array(datapoint)
    centroid_points = array(centroid_points)
    diff = datapoint - centroid_points
    diffsq = diff**2

    distances = (diffsq.sum(axis = 1))**0.5
    # Get the nearest centroid for each instance
    min_dist = min(distances)
    return min_dist

class Quesion11(MRJob):

    def steps(self):
        return [
            MRJobStep(mapper_init = self.mapper_init, mapper=self.mapper,
                      reducer_init=self.reducer_init,
                      reducer=self.reducer,
                      reducer_final=self.reducer_final)
                ]
    #load centroids info from file
    def mapper_init(self):
        self.centroid_points = [[-4.5,0.0], [4.5,0.0], [0.0,4.5]]

    #load data and output the nearest centroid index and data point
    def mapper(self, _, line):
        D = (map(float,line.split(',')))
        ######################## let's weight the input ########################
        norm = sqrt(sum([x**2 for x in D]))
        D = [x/norm for x in D]

        w_dist = MinDist(D, self.centroid_points)/norm
        yield None, (w_dist, 1.0/norm)

    def reducer_init(self):
        self.sum_d = 0
        self.sum_w = 0

    def reducer(self, _, value):
        for v in value:
            d, w = v[0], v[1]
            self.sum_d += d
            self.sum_w += w

    def reducer_final(self):
        yield None, self.sum_d/self.sum_w


if __name__ == '__main__':
    Quesion11.run()
```

```
Overwriting Quesion11.py
```

In [14]:
```
!python Quesion11.py Kmeandata.csv
```

```
using configs in /Users/leiyang/.mrjob.conf
creating tmp directory /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.20160303
.013448.608016
MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.20160303.013448.6080
16/step-0-mapper_part-00000
MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
Counters from step 1:
  (no counters found)
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.20160303.013448.6080
16/step-0-mapper-sorted
> sort /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.20160303.013448.608016/s
tep-0-mapper_part-00000
writing to /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.20160303.013448.6080
16/step-0-reducer_part-00000
```

```
MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
MRJobStep has been renamed to MRStep. The old name will be removed in v0.5.0.
Counters from step 1:
  (no counters found)
Moving /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.20160303.013448.608016/s
tep-0-reducer_part-00000 -> /var/folders/tx/5ldq67q511q8wqwqkvptnxd00000gn/T/Quesion11.leiyang.201
60303.013448.608016/output/part-00000
```