In [44]:
```
%%file q1.py
from mrjob.job import MRJob
from mrjob.step import MRStep
from math import sqrt

from mrjob.job import MRJob
from mrjob.step import MRStep
from math import sqrt

class MRKMeans(MRJob):

    def dist_vec(self, v1, v2):
        # Calculate the distance between two vectors (in two dimensions)
        return sqrt((v2[0] - v1[0]) ** 2 + (v2[1] - v1[1]) ** 2)

    def configure_args(self):
        super(MRKMeans, self).configure_args()
        self.add_file_arg('--centroids-file', dest='centroids_file', help='Path to centroids file')

    def get_centroids(self):
        centroids = []
        with open(self.options.centroids_file, 'r') as f:
            for line in f:
                x, y = map(float, line.strip().split(','))
                centroids.append([x, y])
        return centroids

    def mapper(self, _, lines):
        centroids = self.get_centroids()
        for line in lines.split('\n'):
            x, y = map(float, line.strip().split(','))
            point = [x, y]
            min_dist = float('inf')
            class_index = 0
            for i, centroid in enumerate(centroids):
                dist = self.dist_vec(point, centroid)
                if dist < min_dist:
                    min_dist = dist
                    class_index = i
            yield class_index, (point,1)

    def combiner(self, class_index, points):
        count=0
        sum_x = sum_y = 0.0
        for point in points:
            count = count+point[1]
            sum_x += point[0][0]
            sum_y += point[0][1]
        yield class_index, ((sum_x ,sum_y),count)

    def reducer(self, class_index, points):
        count=0
        sum_x = sum_y = 0.0
        for point in points:
            count = count+point[1]
            sum_x += point[0][0]
            sum_y += point[0][1]
        yield class_index, ((sum_x/count ,sum_y/count))

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

```
Overwriting q1.py
```

In [45]:
```
!python q1.py --centroids-file centroids.txt sampleTest.txt
```

```
0       [1.5,2.5]
1       [3.5,4.5]
2       [7.5,8.5]

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240314.183343.824286
Running step 1 of 1...
job output is in C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240314.183343.824286\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240314.183343.824286\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240314.183343.824286...
```

In [ ]: