Data Intensive Programming 2018

Assignment

Task completed: 1,2,3,4,5,6

Pino Surace (262767): task 1,2,3,4

Khoa Nguyen (272580): task 5, 6

Mikko Saari (245759): supports all the tasks and structures the code

Task 3:

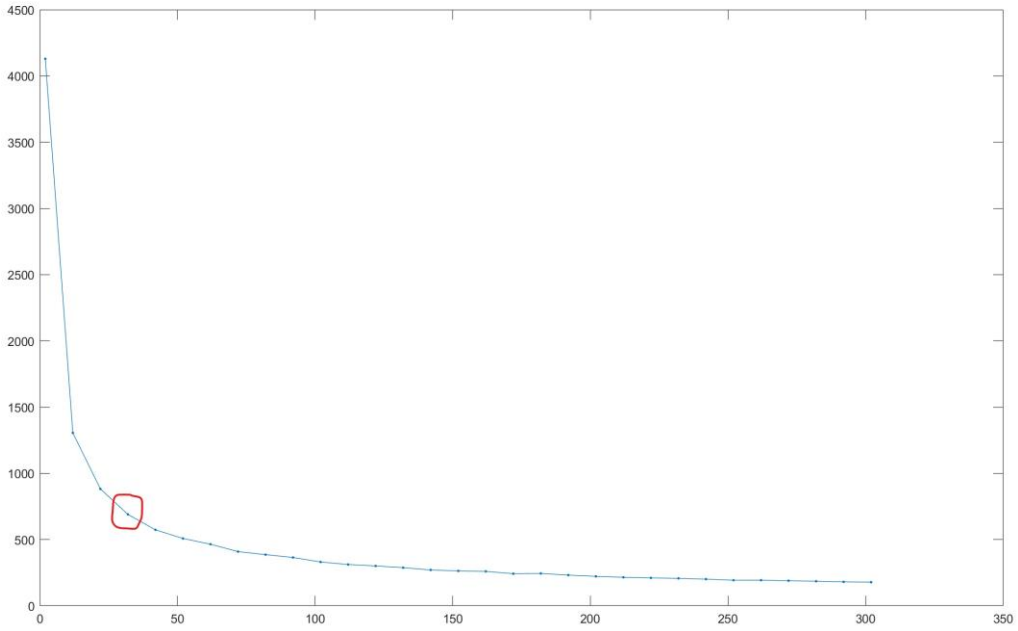
```
//Get streaming data
val  streamingDF: DataFrame =  spark.readStream
                                    .format("csv")
                                    .option("sep", ",")
                                    .option("header", "true")
                                    .schema(staticSchema)
                                    .load("streamingData/*.csv")


  //run k-means with 10 clusters
 val vectorAssembler  = new VectorAssembler()
                              .setInputCols(Array("X","Y"))
                              .setOutputCol("features")
   val transformationPipeline = new Pipeline().setStages(Array(vectorAssembler))
   val coordinates : DataFrame = data.select("X", "Y")
   val pipeLine = transformationPipeline.fit(coordinates)
   val transformedTraining = pipeLine.transform(coordinates)
   val kmeans = new KMeans().setK(10).setSeed(1L)
   val kmModel = kmeans.fit(transformedTraining)


// print results on console
  kmModel.summary.predictions.writeStream
   .format("console")
   .queryName("k means")
   .outputMode("complete")
```

.start()

Task4:

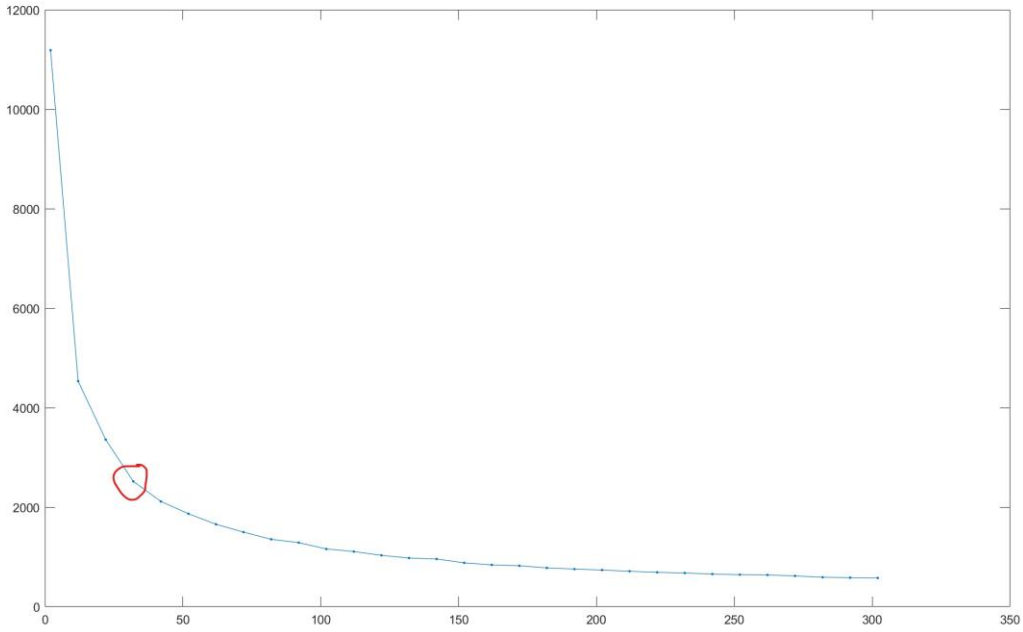
Two dimensions:

The elbow point is not clearly showed in the picture because we are working with real data. It seems that it could be about k = 32.



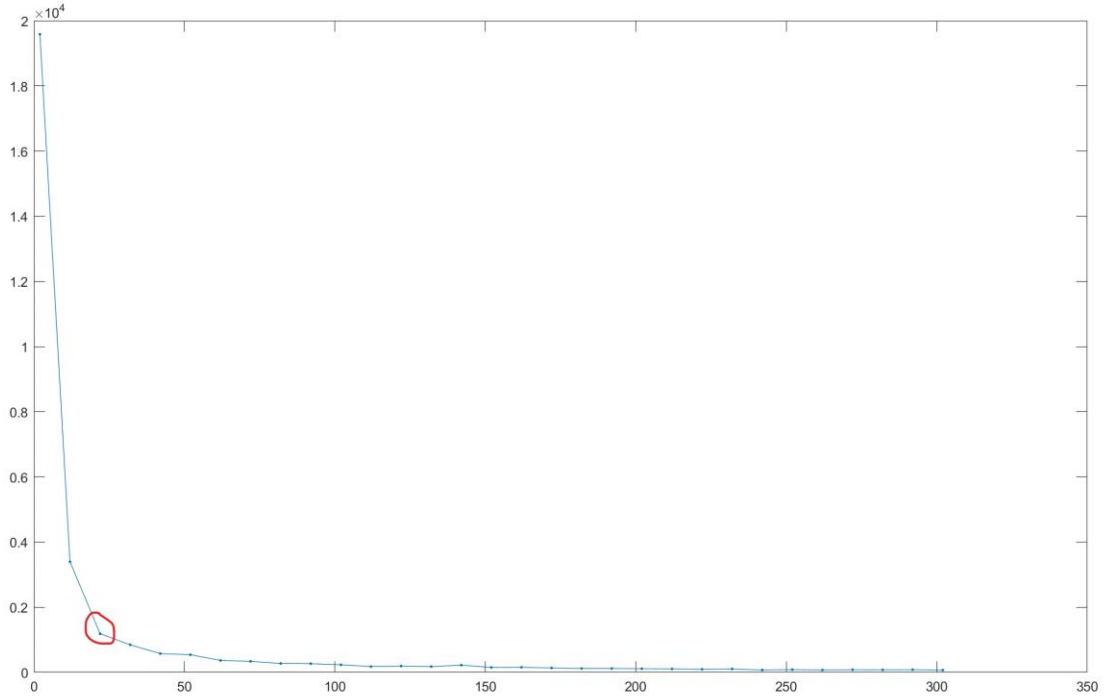
Three dimensions:

The elbow point is not clearly showed in the picture because we are working with real data. It seems that it could be about k = 32.



Task 6:

The elbow point in the algorithm made by us seems to be more evident and exactly to be equal to k = 22.


×10^4
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
50
100
150
200
250
300
350