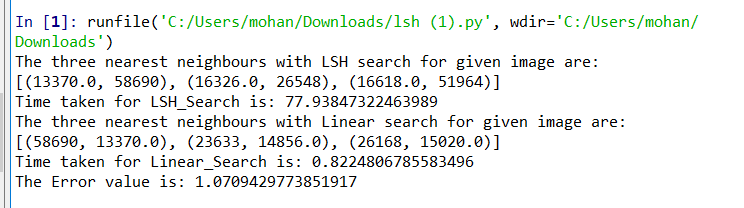
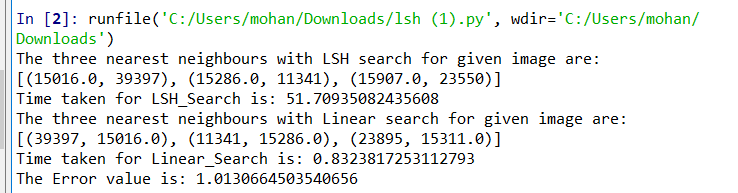
1. **For each of the image patches(from 100 to 1000), the top 3 nearest neighbors using both LSH and linear search**

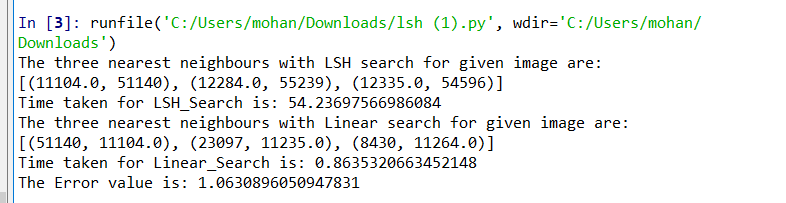
The top three nearest neighbor for 99th index( with k=24,L=10 ) – 100th row:



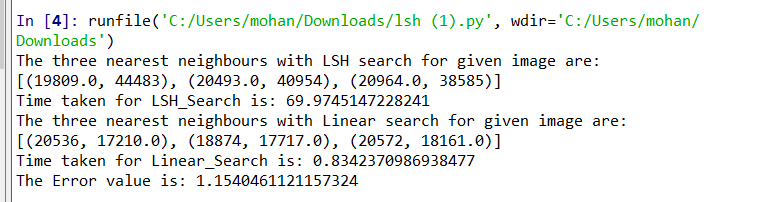
199th index – 200th row:



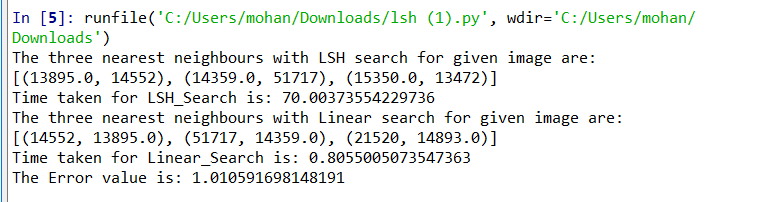
299th index – 300th row:



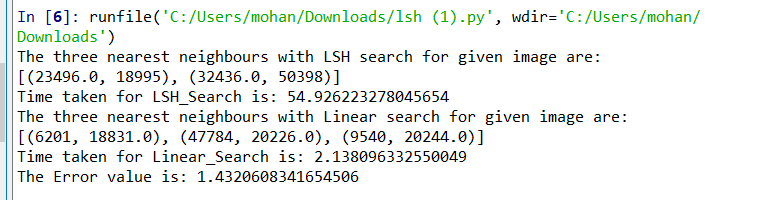
399th Index - 400th row:



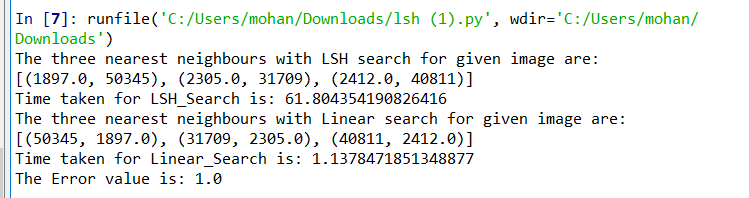
499th index – 500th row:



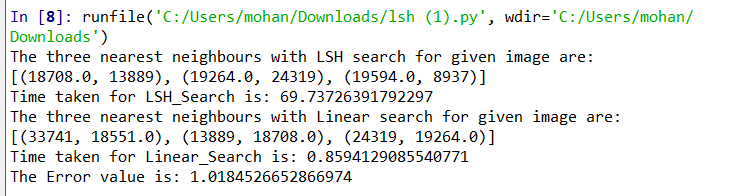
599th index – 600th row:



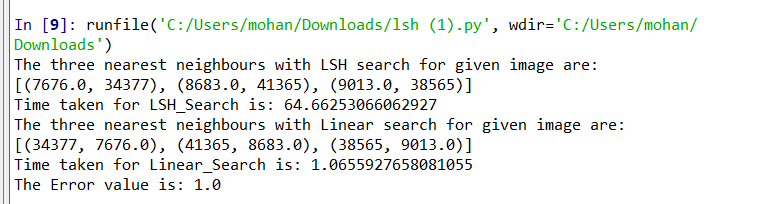
699th index- 700th row:



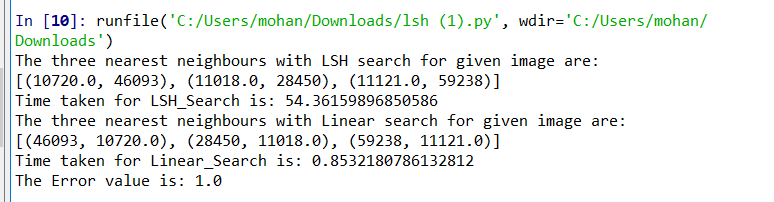
799th index- 800th row:



899th index – 900th row:



999th index – 1000th row:



Average search time for LSH = (sum of time taken to find 3 nearest neighbors for 100th to 1000th row)/10

= 62.935 seconds

* Average search time using LSH is 50.485 seconds

Average search time for Linear search = (sum of time taken for linear search for 100th to 1000th )/10

= 1.039 seconds

* Average search time using Linear search is 1.083 seconds

1. **Error Measure computation :**

Error Measure computation =1/10(Sum of error value of distances from 100th row to 1000th row)

= 1/10(10.7623)

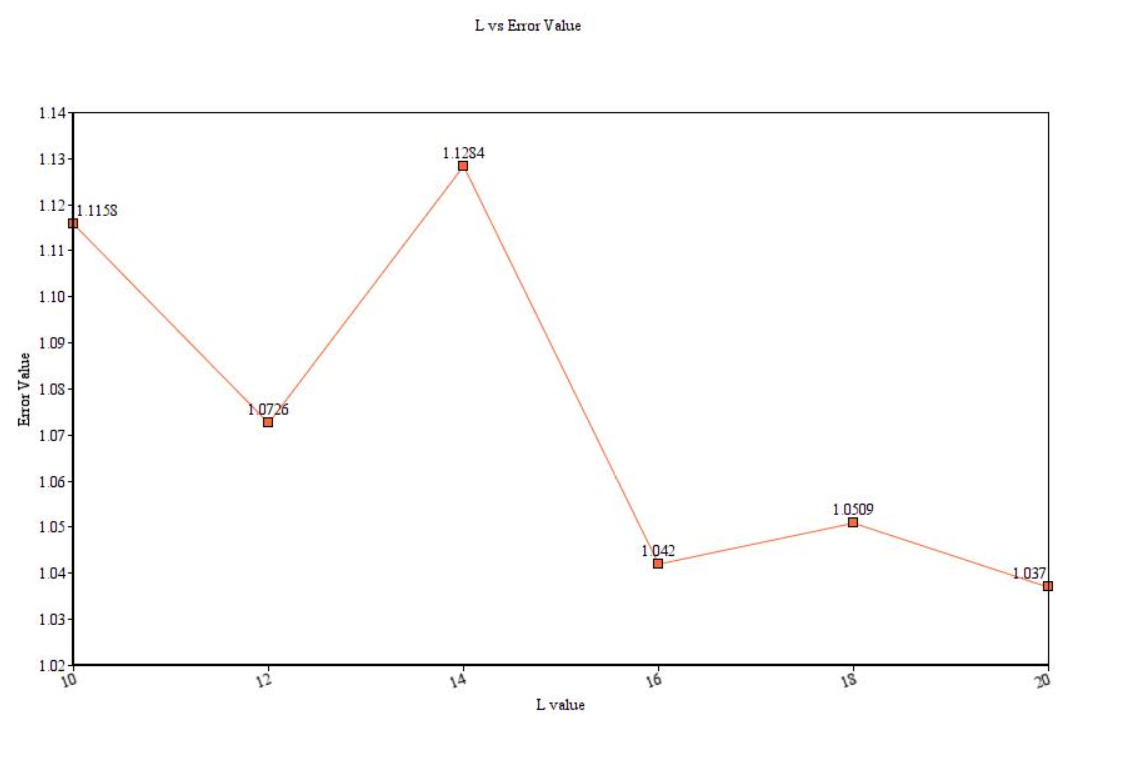
= 1.0762

* Error Measure value with the provided formula is 1.0762

The error for LSH and linear is displayed in the above screenshots as error value.

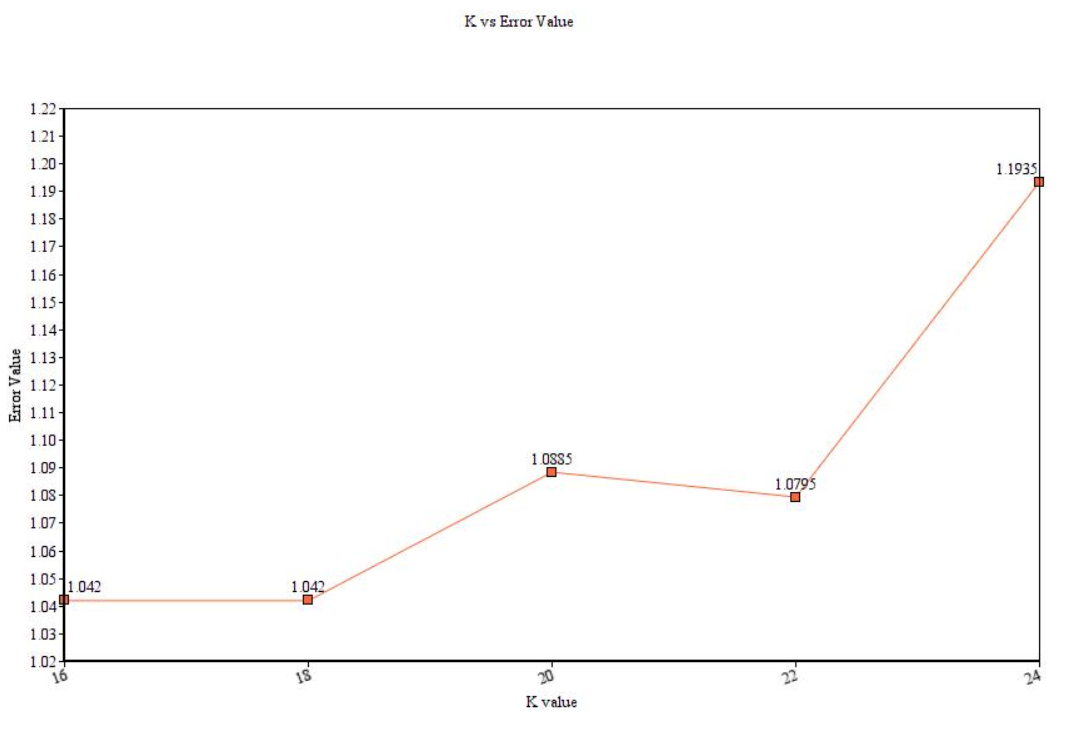
1. **Plotting of Error value vs K and Error value vs L**

**For K=24, Error value vs L plot will be as below :**

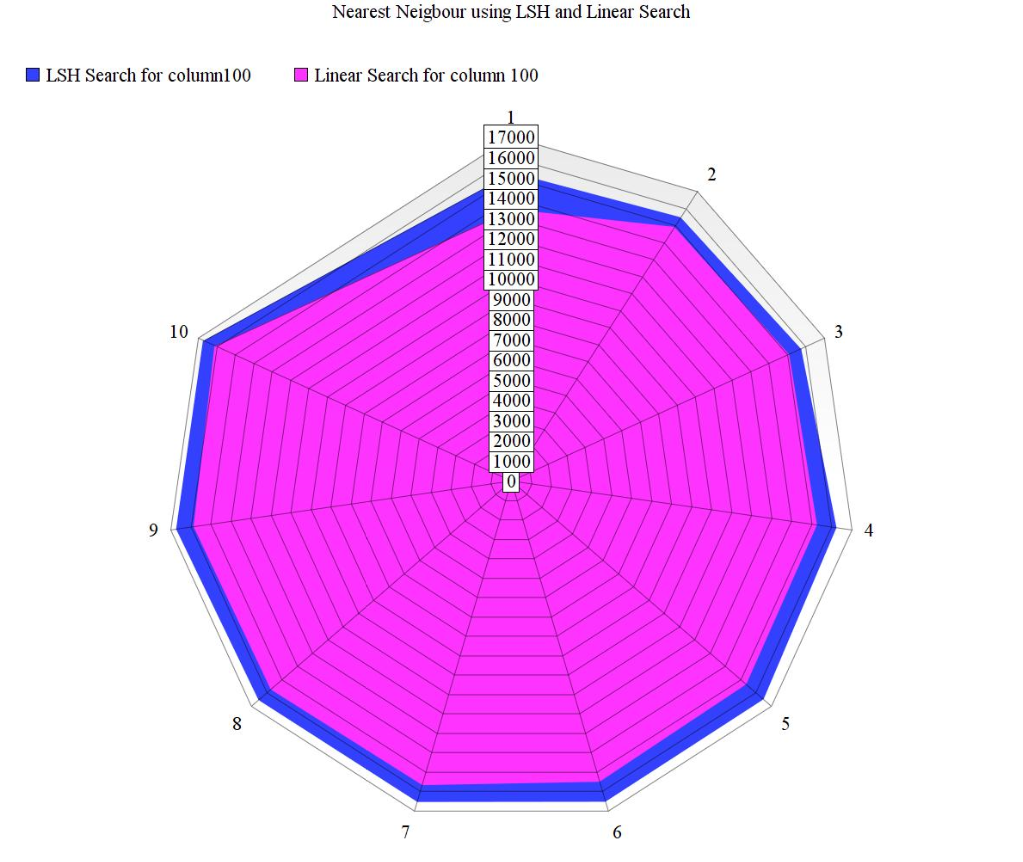


From the above plot it is understood that, if length of hash function per image or row increases(L value), error value is showing decreasing trend. As the number of hash functions increasing, it means we are increasing the probability of finding a candidate pair. It has more opportunity to find the similar image patches.

**For L=10, K vs Error value plot will be as below:**



From the above plot, it is understood that, if the length of hash vector (K value) increases, error value shows increasing trend. It is evident that if K value increases, the probability of finding similar candidate pairs decreases and increases the error value subsequently.



From the above chart , 0 represents the image patch of column 100 and the points 1 to 10 are the nearest and distance from the image patch.

Pink color data represents – Ten nearest Neighbors using linear search

Blue color data represents – Ten nearest Neighbors using LSH search

From the above visual comparison, LSH will provide closest information and will be more useful in finding similar images if the data size is very big.

**Major Problems encountered in this assignment :**

1. **Pickling Error :**

I got this error saying that attempting reference of spark context from a broadcast variable, action or transformation – SPARK 5063

The above error occurred when calling the function to find the distance between two vectors (candidate pair and query index). Hence, I have implemented the distance function inside of LSH search function which resolved the issue.

1. **Python Worker failed to connect back:**

This error caused when I am attempting to get the nearest neighbors with different L and K values in a loop , by closing the current spark session and re-connecting again. I found from stackoverflow.com, that new version of Spark has issues while re-connecting to spark session. Due to this , I am unable to get the top three nearest neighbors and error value in single run.

Hence, I re-ran the script manually changing the values of L and K and plotted the graph.