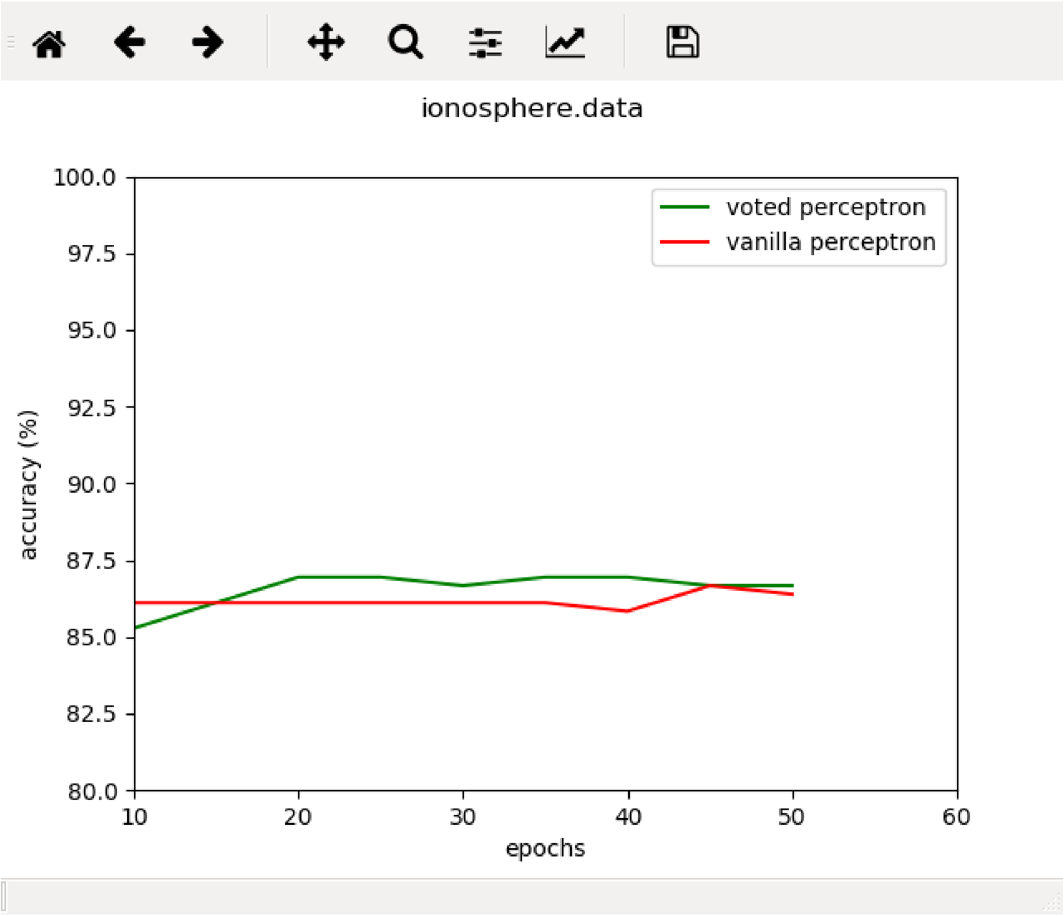
SMAI Assignment 1 Report

Problem 1 : Voted Perceptron

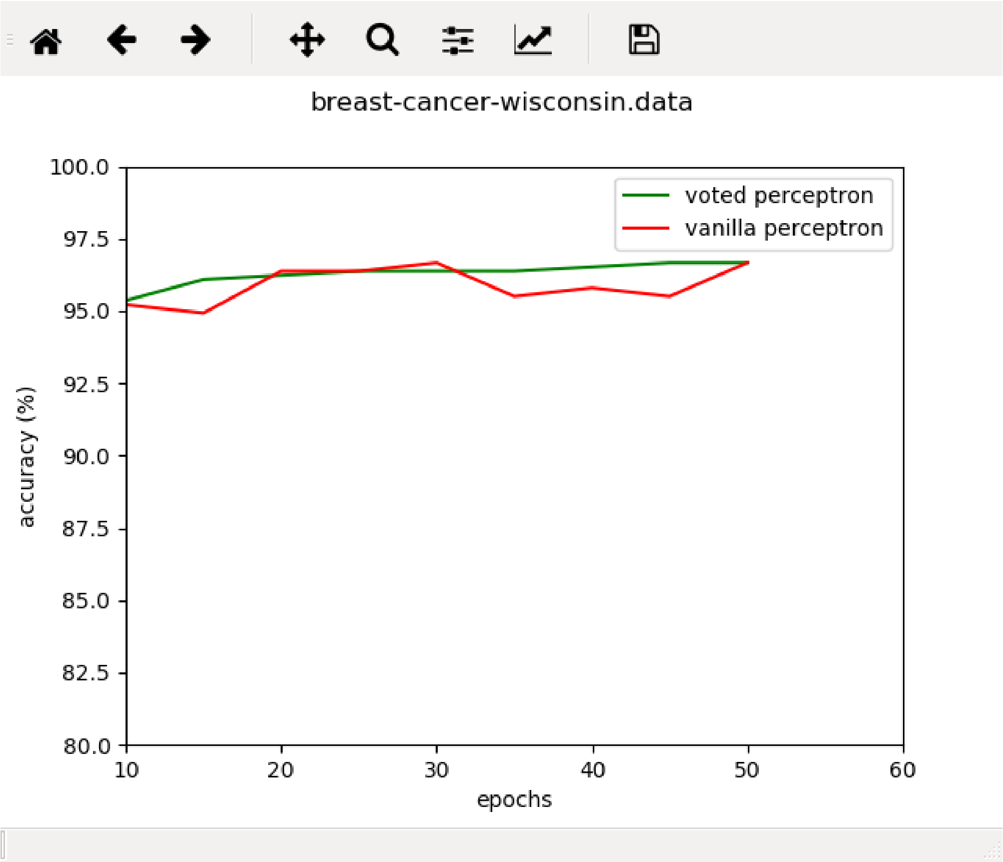
Ionosphere Data (10-fold cross validation, step_size = 1)

vannila perceptron		voted perceptron	
Epochs	Accuracy	Epochs	Accuracy
10	0.86111111111111	10	0.8527777777777776
15	0.86111111111111	15	0.86111111111111
20	0.86111111111111	20	0.8694444444444445
25	0.86111111111111	25	0.8694444444444445
30	0.86111111111111	30	0.8666666666666666
35	0.86111111111111	35	0.8694444444444445
40	0.8583333333333332	40	0.8694444444444445
45	0.8666666666666668	45	0.8666666666666666
50	0.8638888888888889	50	0.8666666666666666



Breast Cancer Wisconsin Data (10-fold cross validation, step_size = 1)

vannila perceptron		voted perceptron	
Epochs	Accuracy	Epochs	Accuracy
10	0.9521739130434785	10	0.9536231884057973
15	0.9492753623188406	15	0.9608695652173914
20	0.963768115942029	20	0.9623188405797102
25	0.963768115942029	25	0.9637681159420289
30	0.9666666666666668	30	0.9637681159420289
35	0.955072463768116	35	0.9637681159420289
40	0.9579710144927536	40	0.9652173913043478
45	0.955072463768116	45	0.9666666666666668
50	0.9666666666666668	50	0.9666666666666668

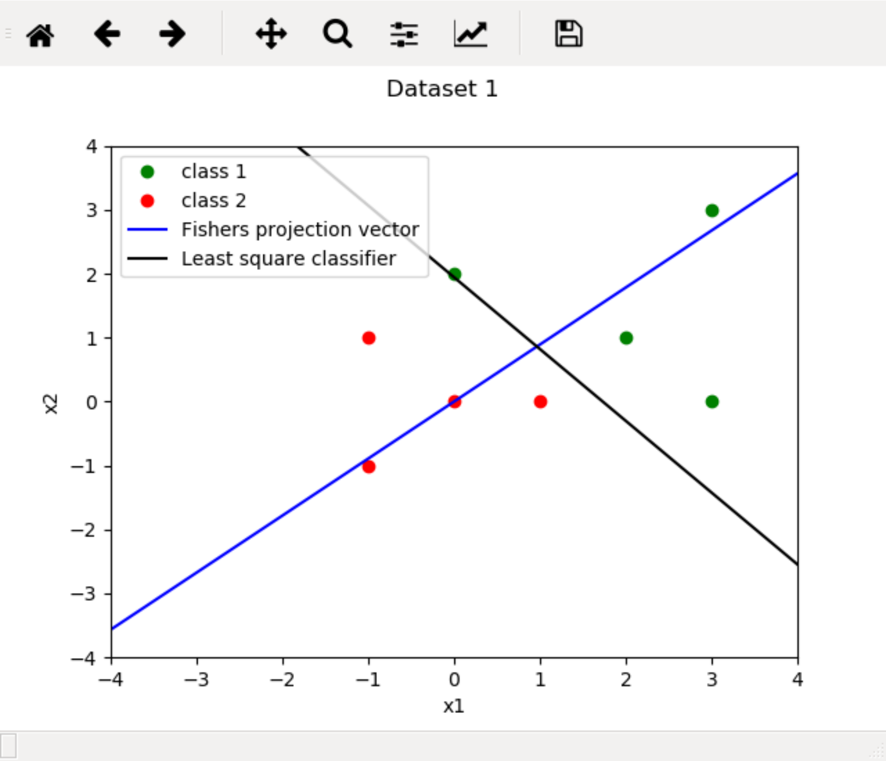


Voted vs Vanilla approach:

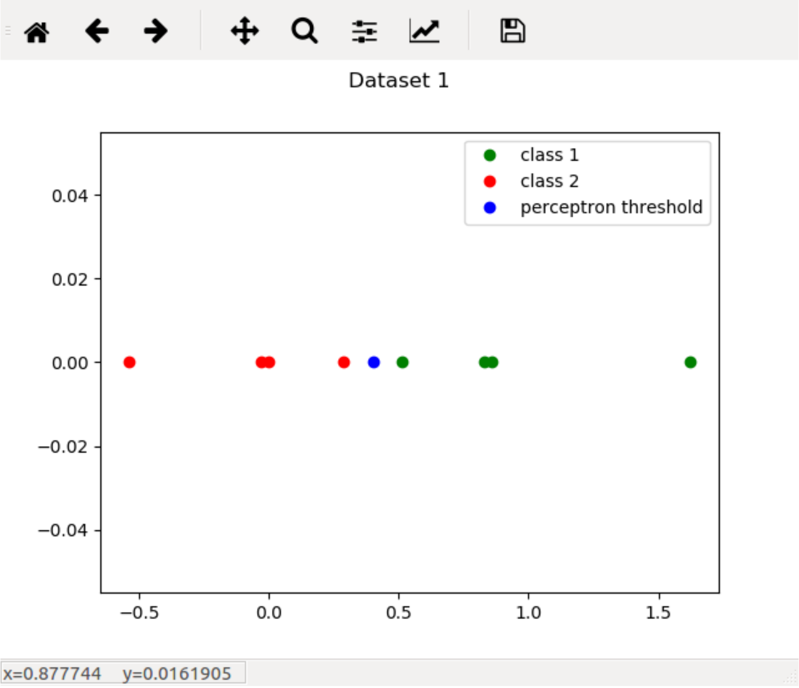
As we can observe from the plot, voted perceptron at most of the times gives higher accuracy than that of the vannilla perceptron, reason being we store the votes of the weight vector. Hence outliers in data does not have much effect on voted approach, where as vanilla might not give satisfactory results in the same case.

Problem 2: Least Square Approach

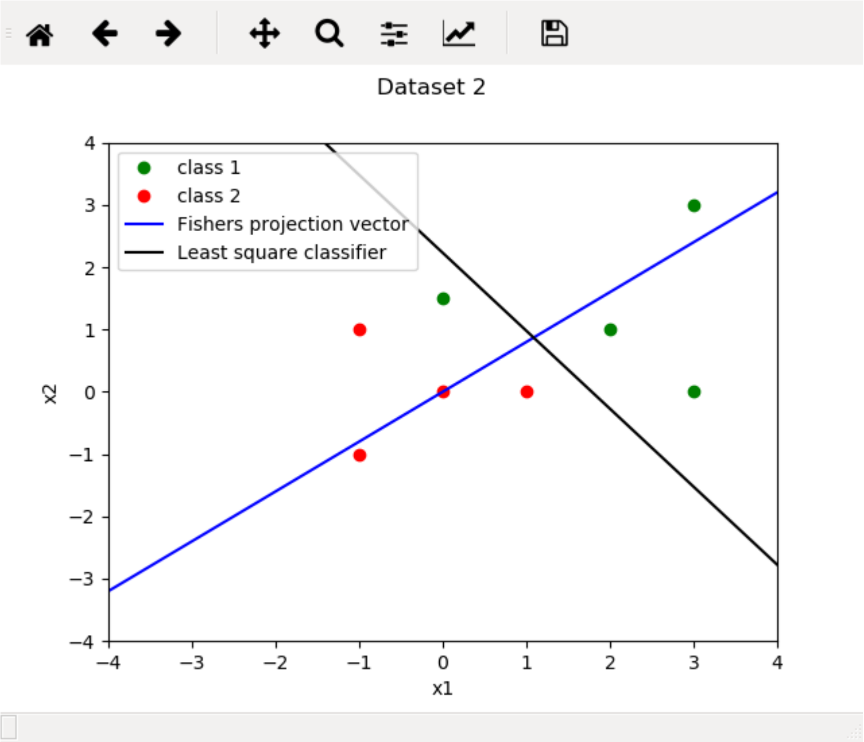
Dataset 1



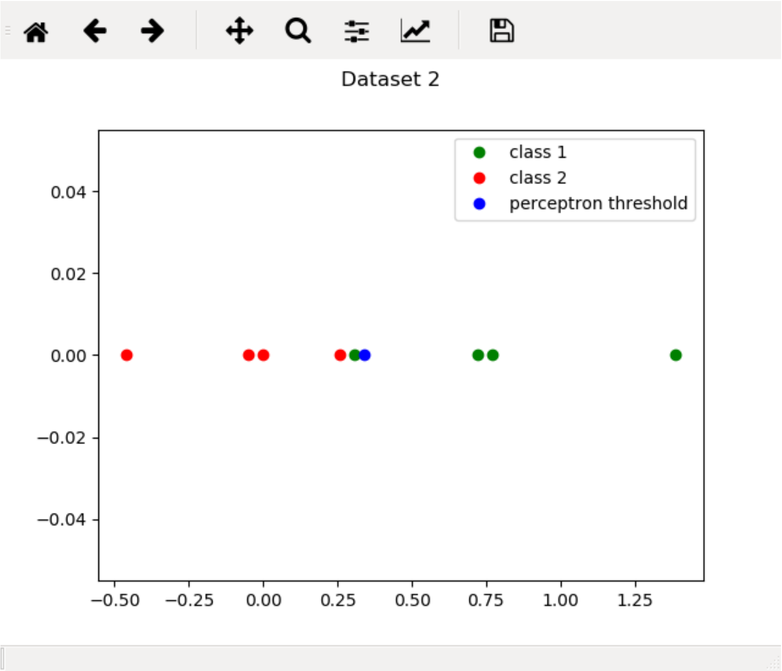
By projecting the dataset on to the Fishers vector, we convert the dataset into set of 1D points, then applied perceptron algorithm (epochs=1, step=0.3) to find the threshold, which is shown below.



Dataset 2



By projecting the dataset on to the Fishers vector, we convert the dataset into set of 1D points, then applied perceptron algorithm (epochs=1, step=0.3) to find the threshold, which is shown below.



Fishers LDA vs MSE classifier:

Fishers LDA method gives us the direction in which the interclass variance is maximum and intraclass variance is minimum, thus by projecting the data points onto the vector gives us maximum class seperability. This is one of the major advantage of F-LDA when compared to PCA(which gives us the direction of max variance of the dataset).

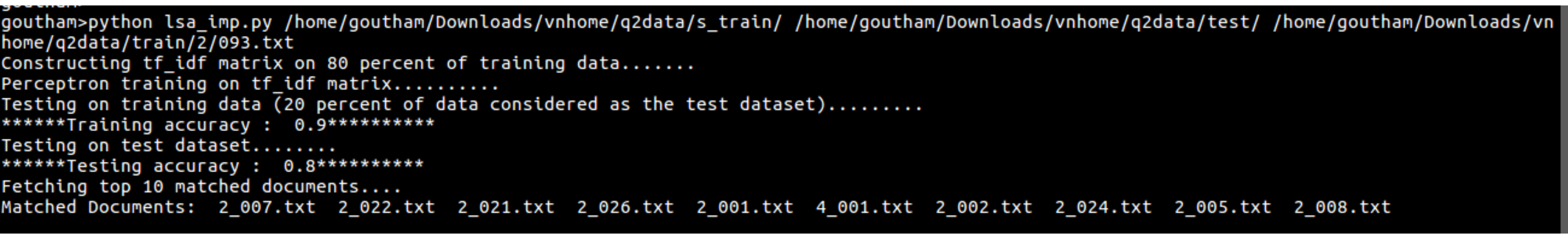
MSE (Minimum Square error) is the process of solving the set of equations i.e Xw = Y using linear algebra, where (X-> dataset, Y->class label, w->solution space), by minimizing the cost function pow(|| Xw-Y||,2).

Observations:

- In case of F-LDA and MSE, overlapping data (dataset-2) gives poor results.
- In non-overlapping data(dataset-1), F-LDA along with perceptron gives a satisfactory result, while MSE barely classifies the data.

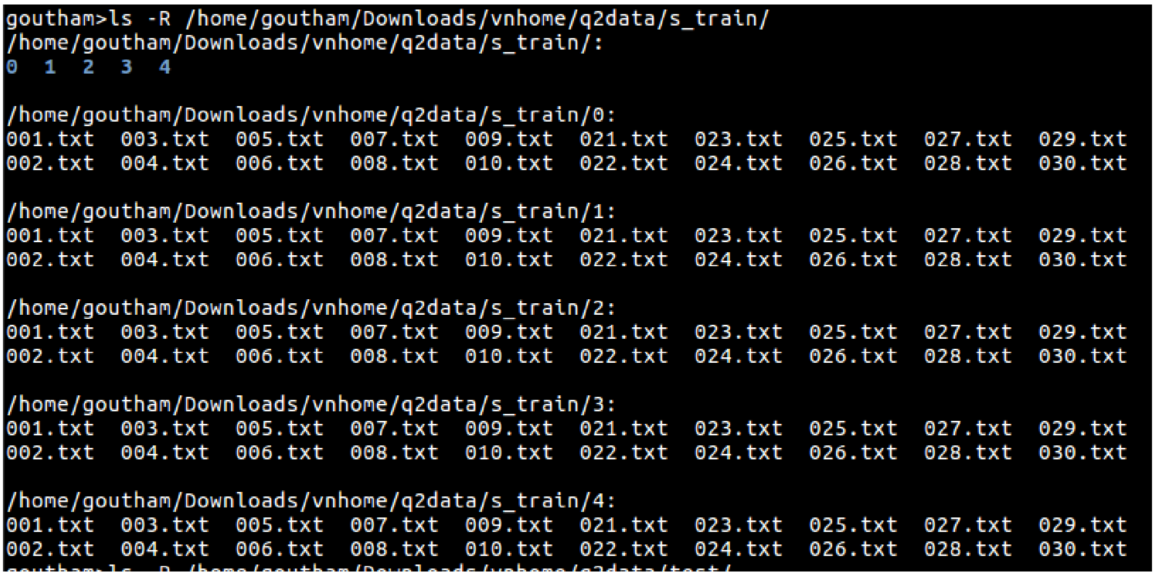
Problem 3: Latent Semantic Analysis

The below screenshot shows the
1) accuracy on train dataset (20% of train data)
2) testing accuracy (on the test dataset)
3) query matching results (of doc belonging to class 2 file 93.txt)
4) Unable to verify dimensionality reduction results as the computation time for each svd was taking too long.



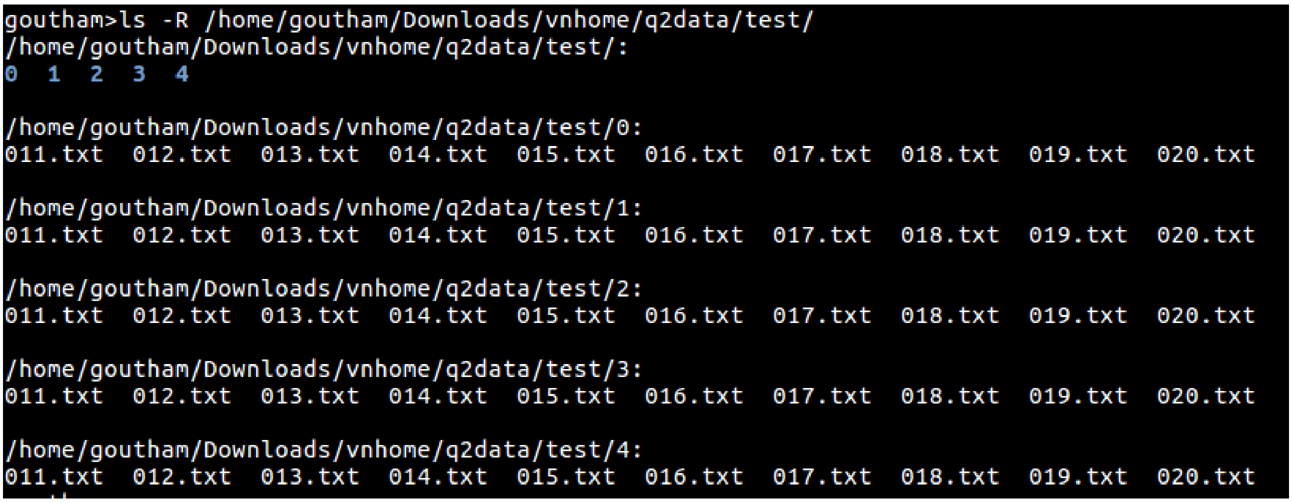
```
goutham>python lsa_imp.py /home/goutham/Downloads/vnhome/q2data/s_train/ /home/goutham/Downloads/vnhome/q2data/test/ /home/goutham/Downloads/vn
home/q2data/train/2/093.txt
Constructing tf_idf matrix on 80 percent of training data.......
Perceptron training on tf_idf matrix..........
Testing on training data (20 percent of data considered as the test dataset).........
******Training accuracy :  0.9**********
Testing on test dataset........
******Testing accuracy :  0.8**********
Fetching top 10 matched documents....
Matched Documents:  2_007.txt  2_022.txt  2_021.txt  2_026.txt  2_001.txt  4_001.txt  2_002.txt  2_024.txt  2_005.txt  2_008.txt
```

Training data:



```
goutham>ls -R /home/goutham/Downloads/vnhome/q2data/s_train/
/home/goutham/Downloads/vnhome/q2data/s_train/:
0  1  2  3  4

/home/goutham/Downloads/vnhome/q2data/s_train/0:
001.txt  003.txt  005.txt  007.txt  009.txt  021.txt  023.txt  025.txt  027.txt  029.txt
002.txt  004.txt  006.txt  008.txt  010.txt  022.txt  024.txt  026.txt  028.txt  030.txt

/home/goutham/Downloads/vnhome/q2data/s_train/1:
001.txt  003.txt  005.txt  007.txt  009.txt  021.txt  023.txt  025.txt  027.txt  029.txt
002.txt  004.txt  006.txt  008.txt  010.txt  022.txt  024.txt  026.txt  028.txt  030.txt

/home/goutham/Downloads/vnhome/q2data/s_train/2:
001.txt  003.txt  005.txt  007.txt  009.txt  021.txt  023.txt  025.txt  027.txt  029.txt
002.txt  004.txt  006.txt  008.txt  010.txt  022.txt  024.txt  026.txt  028.txt  030.txt

/home/goutham/Downloads/vnhome/q2data/s_train/3:
001.txt  003.txt  005.txt  007.txt  009.txt  021.txt  023.txt  025.txt  027.txt  029.txt
002.txt  004.txt  006.txt  008.txt  010.txt  022.txt  024.txt  026.txt  028.txt  030.txt

/home/goutham/Downloads/vnhome/q2data/s_train/4:
001.txt  003.txt  005.txt  007.txt  009.txt  021.txt  023.txt  025.txt  027.txt  029.txt
002.txt  004.txt  006.txt  008.txt  010.txt  022.txt  024.txt  026.txt  028.txt  030.txt
goutham>ls -R /home/goutham/Downloads/vnhome/q2data/test/
```

Testing data:



```
goutham>ls -R /home/goutham/Downloads/vnhome/q2data/test/
/home/goutham/Downloads/vnhome/q2data/test/:
0  1  2  3  4

/home/goutham/Downloads/vnhome/q2data/test/0:
011.txt  012.txt  013.txt  014.txt  015.txt  016.txt  017.txt  018.txt  019.txt  020.txt

/home/goutham/Downloads/vnhome/q2data/test/1:
011.txt  012.txt  013.txt  014.txt  015.txt  016.txt  017.txt  018.txt  019.txt  020.txt

/home/goutham/Downloads/vnhome/q2data/test/2:
011.txt  012.txt  013.txt  014.txt  015.txt  016.txt  017.txt  018.txt  019.txt  020.txt

/home/goutham/Downloads/vnhome/q2data/test/3:
011.txt  012.txt  013.txt  014.txt  015.txt  016.txt  017.txt  018.txt  019.txt  020.txt

/home/goutham/Downloads/vnhome/q2data/test/4:
011.txt  012.txt  013.txt  014.txt  015.txt  016.txt  017.txt  018.txt  019.txt  020.txt
```