CS771- Assignment 1

Shubham Jain - 13683

1 a) Spam Classification
Taking part 1 as the test data the accuracy is: 0.9862
Taking part 2 as the test data the accuracy is: 0.9828
Taking part 3 as the test data the accuracy is: 0.9897
Taking part 4 as the test data the accuracy is: 0.9897
Taking part 5 as the test data the accuracy is: 1.0000
Taking part 6 as the test data the accuracy is: 1.0000
Taking part 7 as the test data the accuracy is: 0.9897
Taking part 8 as the test data the accuracy is: 0.9655
Taking part 9 as the test data the accuracy is: 0.9931
Taking part 10 as the test data the accuracy is: 0.9932

Average Accuracy = 0.98899

1 b) After stop words are removed:
Taking part 1 as the test data the accuracy is: 0.9897
Taking part 2 as the test data the accuracy is: 0.9862
Taking part 3 as the test data the accuracy is: 0.9862
Taking part 4 as the test data the accuracy is: 0.9931
Taking part 5 as the test data the accuracy is: 1.0000
Taking part 6 as the test data the accuracy is: 1.0000
Taking part 7 as the test data the accuracy is: 0.9931
Taking part 8 as the test data the accuracy is: 0.9759
Taking part 9 as the test data the accuracy is: 0.9828
Taking part 10 as the test data the accuracy is: 0.9932

Average Accuracy = 0.99002

1 c) After Lemmatizing the email:
Taking part 1 as the test data the accuracy is: 0.9862
Taking part 2 as the test data the accuracy is: 0.9862
Taking part 3 as the test data the accuracy is: 0.9862
Taking part 4 as the test data the accuracy is: 0.9931
Taking part 5 as the test data the accuracy is: 0.9966
Taking part 6 as the test data the accuracy is: 0.9966
Taking part 7 as the test data the accuracy is: 0.9862
Taking part 8 as the test data the accuracy is: 0.9690
Taking part 9 as the test data the accuracy is: 0.9759
Taking part 10 as the test data the accuracy is: 0.9932

Average Accuracy = 0.98692

2 Different Bag of Words representations

a) Binary bag of words representation:
Taking part 1 as the test data the accuracy is: 0.9897
Taking part 2 as the test data the accuracy is: 0.9966
Taking part 3 as the test data the accuracy is: 0.9897
Taking part 4 as the test data the accuracy is: 0.9931
Taking part 5 as the test data the accuracy is: 0.9966
Taking part 6 as the test data the accuracy is: 0.9966
Taking part 7 as the test data the accuracy is: 1.0000
Taking part 8 as the test data the accuracy is: 0.9897
Taking part 9 as the test data the accuracy is: 0.9897
Taking part 10 as the test data the accuracy is: 0.9932

Average Accuracy = 0.99349

b) Term frequency bag of words representation:
Taking part 1 as the test data the accuracy is: 0.9793
Taking part 2 as the test data the accuracy is: 0.9862
Taking part 3 as the test data the accuracy is: 0.9897
Taking part 4 as the test data the accuracy is: 0.9793
Taking part 5 as the test data the accuracy is: 0.9828
Taking part 6 as the test data the accuracy is: 0.9897
Taking part 7 as the test data the accuracy is: 0.9897
Taking part 8 as the test data the accuracy is: 0.9862
Taking part 9 as the test data the accuracy is: 0.9931
Taking part 10 as the test data the accuracy is: 0.9760

Average Accuracy = 0.9852

c) Term frequency - Inverse document frequency representation:
Taking part 1 as the test data the accuracy is: 0.9655
Taking part 2 as the test data the accuracy is: 0.9759
Taking part 3 as the test data the accuracy is: 0.9793
Taking part 4 as the test data the accuracy is: 0.9793
Taking part 5 as the test data the accuracy is: 1.0000
Taking part 6 as the test data the accuracy is: 0.9690
Taking part 7 as the test data the accuracy is: 0.9793
Taking part 8 as the test data the accuracy is: 0.9690
Taking part 9 as the test data the accuracy is: 0.9793
Taking part 10 as the test data the accuracy is: 0.9932

Average Accuracy = 0.97898

Q 3) kNN Classifier for MNIST dataset:
Values of k = {1,2,3,4}
Metrics used: { ‘cityblock’, ‘euclidean’, ‘L3’}

Metric	k	Average accuracy
Cityblock	1	0.9631
Euclidean	1	0.9691
L3	1	0.9717
Cityblock	2	0.954
Euclidean	2	0.9627
L3	2	0.9668
Cityblock	3	0.9633
Euclidean	3	0.9705
L3	3	0.9718
Cityblock	4	0.9607
Euclidean	4	0.9682
L3	4	0.9712

Metric	k	Average accuracy
Cityblock	1	0.9631
Cityblock	2	0.954
Cityblock	3	0.9633
Cityblock	4	0.9607
Euclidean	1	0.9691
Euclidean	2	0.9627
Euclidean	3	0.9705
Euclidean	4	0.9682
L3	1	0.9717
L3	2	0.9668
L3	3	0.9718
L3	4	0.9712

Detailed Data:

metic used l1 number of neighbours 1

Average Accuracy = 0.9631

	precision	recall	f1-score	support
0	0.97	0.99	0.98	980
1	0.95	0.99	0.97	1135
2	0.98	0.96	0.97	1032
3	0.95	0.96	0.95	1010
4	0.97	0.95	0.96	982
5	0.94	0.95	0.95	892
6	0.98	0.98	0.98	958
7	0.96	0.96	0.96	1028
8	0.99	0.92	0.95	974
9	0.94	0.96	0.95	1009
avg / total	0.96	0.96	0.96	10000

metic used l2 number of neighbours 1

Average Accuracy = 0.9691

	precision	recall	f1-score	support
0	0.98	0.99	0.99	980
1	0.97	0.99	0.98	1135
2	0.98	0.96	0.97	1032
3	0.96	0.96	0.96	1010
4	0.97	0.96	0.97	982
5	0.95	0.96	0.96	892
6	0.98	0.99	0.98	958
7	0.96	0.96	0.96	1028
8	0.98	0.94	0.96	974
9	0.96	0.96	0.96	1009
avg / total	0.97	0.97	0.97	10000

metic used l3 number of neighbours 1

Average Accuracy = 0.9717

	precision	recall	f1-score	support
0	0.98	0.99	0.99	980
1	0.98	0.99	0.99	1135
2	0.98	0.97	0.98	1032
3	0.97	0.96	0.96	1010
4	0.97	0.96	0.97	982
5	0.96	0.97	0.96	892
6	0.98	0.99	0.98	958
7	0.96	0.97	0.96	1028
8	0.98	0.95	0.97	974
9	0.96	0.96	0.96	1009
avg / total	0.97	0.97	0.97	10000

metic used l1 number of neighbours 2
Average Accuracy = 0.954

	precision	recall	f1-score	support
0	0.95	1.00	0.97	980
1	0.92	1.00	0.96	1135
2	0.97	0.95	0.96	1032
3	0.92	0.97	0.95	1010
4	0.96	0.97	0.96	982
5	0.94	0.95	0.94	892
6	0.99	0.97	0.98	958
7	0.93	0.95	0.94	1028
8	0.99	0.87	0.93	974
9	0.98	0.91	0.94	1009
avg / total	0.96	0.95	0.95	10000

metic used l2 number of neighbours 2
Average Accuracy = 0.9627

	precision	recall	f1-score	support
0	0.96	1.00	0.98	980
1	0.95	1.00	0.97	1135
2	0.97	0.96	0.97	1032
3	0.94	0.97	0.95	1010
4	0.96	0.98	0.97	982
5	0.95	0.95	0.95	892
6	0.99	0.98	0.99	958
7	0.95	0.95	0.95	1028
8	0.99	0.90	0.94	974
9	0.98	0.93	0.95	1009
avg / total	0.96	0.96	0.96	10000

metic used l3 number of neighbours 2
Average Accuracy = 0.9668

	precision	recall	f1-score	support
0	0.96	0.99	0.98	980
1	0.96	1.00	0.98	1135
2	0.97	0.97	0.97	1032
3	0.94	0.97	0.96	1010
4	0.96	0.98	0.97	982
5	0.96	0.96	0.96	892
6	0.99	0.98	0.98	958
7	0.96	0.96	0.96	1028
8	0.99	0.92	0.96	974
9	0.98	0.93	0.96	1009
avg / total	0.97	0.97	0.97	10000

metic used l1 number of neighbours 3
Average Accuracy = 0.9633

	precision	recall	f1-score	support
0	0.96	0.99	0.98	980
1	0.94	1.00	0.97	1135
2	0.98	0.95	0.97	1032
3	0.96	0.96	0.96	1010
4	0.97	0.95	0.96	982
5	0.96	0.96	0.96	892
6	0.98	0.98	0.98	958
7	0.95	0.96	0.95	1028
8	0.99	0.92	0.95	974
9	0.95	0.95	0.95	1009
avg / total	0.96	0.96	0.96	10000

metic used l2 number of neighbours 3
Average Accuracy = 0.9705

	precision	recall	f1-score	support
0	0.97	0.99	0.98	980
1	0.96	1.00	0.98	1135
2	0.98	0.97	0.97	1032
3	0.96	0.97	0.96	1010
4	0.98	0.97	0.97	982
5	0.97	0.96	0.96	892
6	0.98	0.99	0.98	958
7	0.96	0.96	0.96	1028
8	0.99	0.94	0.96	974
9	0.96	0.96	0.96	1009
avg / total	0.97	0.97	0.97	10000

metic used l3 number of neighbours 3
Average Accuracy = 0.9718

	precision	recall	f1-score	support
0	0.97	0.99	0.98	980
1	0.96	1.00	0.98	1135
2	0.98	0.97	0.98	1032
3	0.96	0.97	0.97	1010
4	0.98	0.97	0.97	982
5	0.97	0.97	0.97	892
6	0.98	0.99	0.98	958
7	0.96	0.96	0.96	1028
8	0.99	0.94	0.97	974
9	0.96	0.96	0.96	1009
avg / total	0.97	0.97	0.97	10000

metic used l1 number of neighbours 4
Average Accuracy = 0.9607

	precision	recall	f1-score	support
0	0.96	0.99	0.98	980
1	0.93	1.00	0.96	1135
2	0.98	0.94	0.96	1032
3	0.95	0.97	0.96	1010
4	0.97	0.96	0.96	982
5	0.95	0.96	0.96	892
6	0.99	0.98	0.98	958
7	0.94	0.95	0.95	1028
8	0.99	0.91	0.95	974
9	0.97	0.94	0.95	1009
avg / total	0.96	0.96	0.96	10000

metic used l2 number of neighbours 4
Average Accuracy = 0.9682

	precision	recall	f1-score	support
0	0.96	1.00	0.98	980
1	0.95	1.00	0.98	1135
2	0.98	0.96	0.97	1032
3	0.96	0.97	0.97	1010
4	0.97	0.97	0.97	982
5	0.96	0.97	0.96	892
6	0.98	0.98	0.98	958
7	0.95	0.96	0.96	1028
8	0.99	0.93	0.96	974
9	0.97	0.95	0.96	1009
avg / total	0.97	0.97	0.97	10000

metic used l3 number of neighbours 4
Average Accuracy = 0.9712

	precision	recall	f1-score	support
0	0.97	0.99	0.98	980
1	0.96	1.00	0.98	1135
2	0.98	0.97	0.97	1032
3	0.96	0.97	0.97	1010
4	0.97	0.97	0.97	982
5	0.96	0.97	0.97	892
6	0.98	0.98	0.98	958
7	0.96	0.97	0.97	1028
8	0.99	0.94	0.96	974
9	0.97	0.95	0.96	1009
avg / total	0.97	0.97	0.97	10000