BITS F464 – MACHINE LEARNING

Assignment I – Email Classifier

Varkeychan Jacob

2017B5A70828P

Aswin Benedict

2019A4PS0579P

Introduction

Programming Language Used: Python

Emails were processed using TF-IDF vectorizer for extracting the features and also finding the significance of each feature in every mail. 1-gram, 2-gram and 3-gram models were used for the feature extraction.

Only top 20% of the extracted features were used to reduce the dimensionality of the problem. Also, the data set was split into training and test data in the ratio of 7:3.

The classification based on these extracted features were done with five different classifiers:

- Support Vector Machines
- K-Nearest Neighbors
- Decision Trees
- Random Forest
- Multi-Layer Perceptron

For SVM, 4 different kernels were tried out:

- Linear Kernel
- Radial Basis Kernel
- Polynomial Kernel
- Sigmoid Kernel

The emails were classified under 6 labels:

- Academics
- Internships/Placements
- Promotions
- News
- Social
- Miscellaneous/Spam

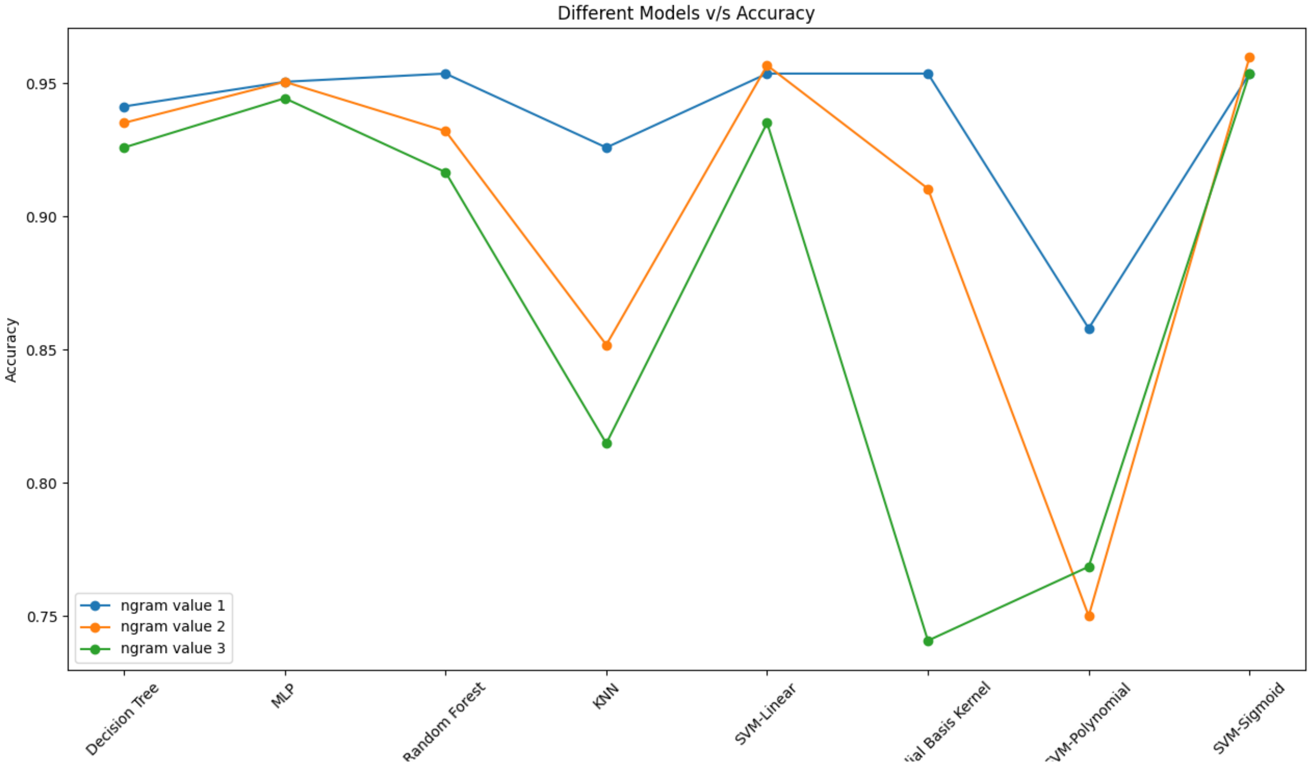
For K-Nearest Neighbors, the number of nearest neighbors were varied from 1 to 50 and were found out that the maximum accuracy was around 6-8 neighbors.

For Decision Trees, the maximum depth was treated as a varying parameter.

For Random Forest, the number of base classifiers were varied.

For Multi-Layer Perceptron, the maximum iterations were varied so that the effect of number of queries over the data could be noted.

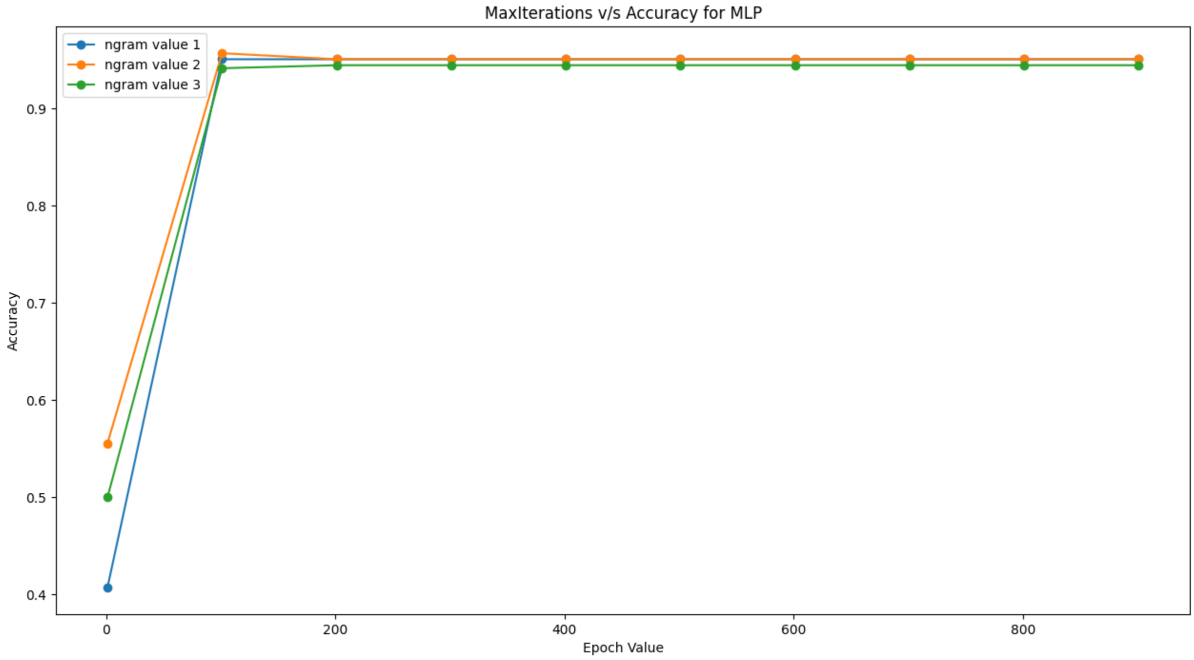
Combined Accuracy for Different Models



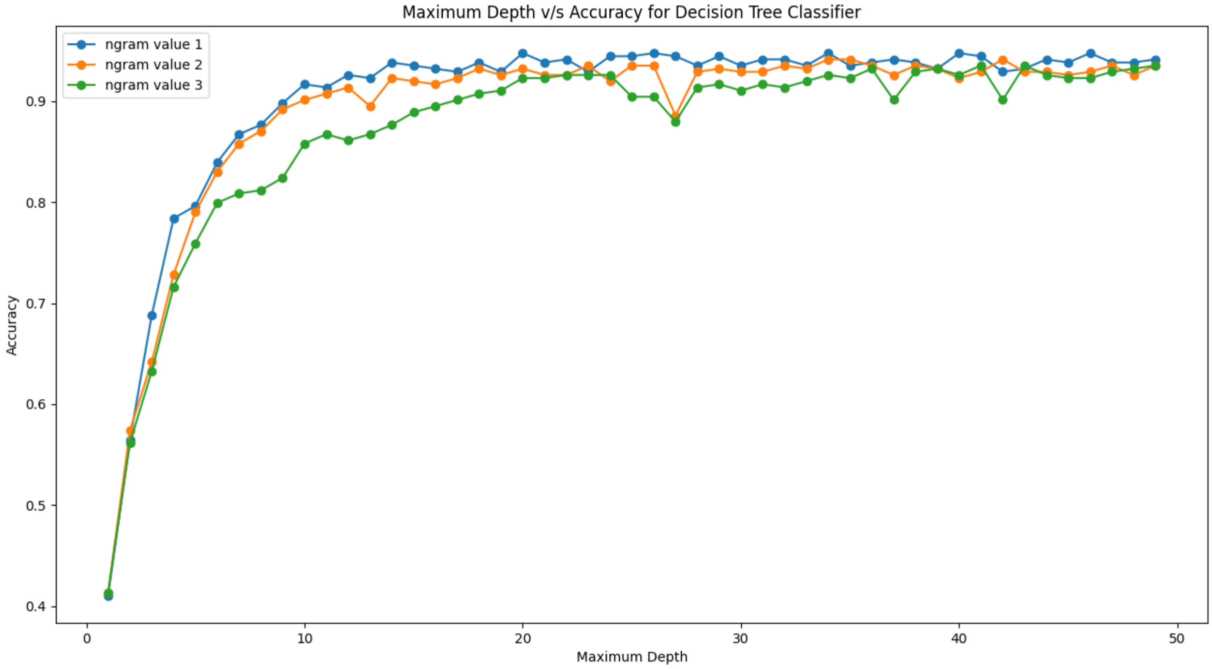
Model	Accuracy
1-gram:	
Decision Trees	93.52%
Random Forest	97.06%
KNN	94.14%
MLP	92.59%
SVM-Linear	95.37%
SVM-Radial Basis Kernel	95.37%
SVM-Polynomial	85.80%
SVM-Sigmoid	97.37%
2-gram:	

Decision Trees	92.28%
Random Forest	95.06%
KNN	92.59%
MLP	85.19%
SVM-Linear	95.68%
SVM-Radial Basis Kernel	91.05%
SVM-Polynomial	75.00%
SVM-Sigmoid	95.99%
3-gram:	
Decision Trees	92.90%
Random Forest	94.44%
KNN	92.28%
MLP	81.48%
SVM-Linear	93.52%
SVM-Radial Basis Kernel	74.07%
SVM-Polynomial	76.85%
SVM-Sigmoid	95.37%

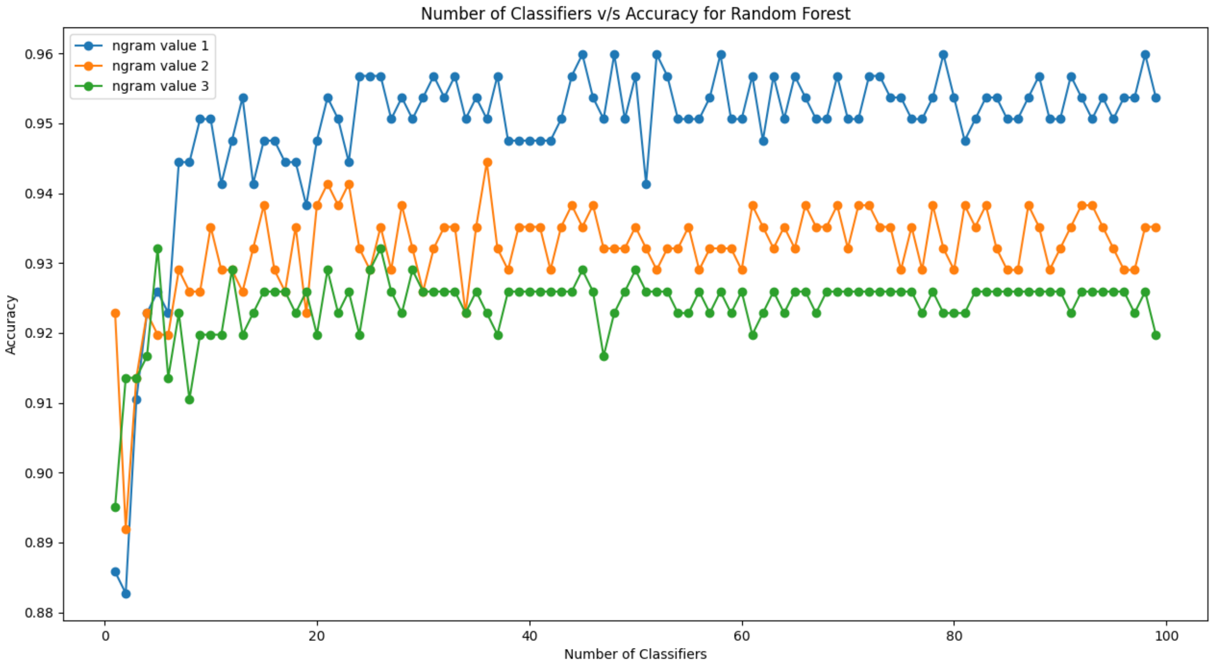
Iterations v/s Accuracy graph for MLP



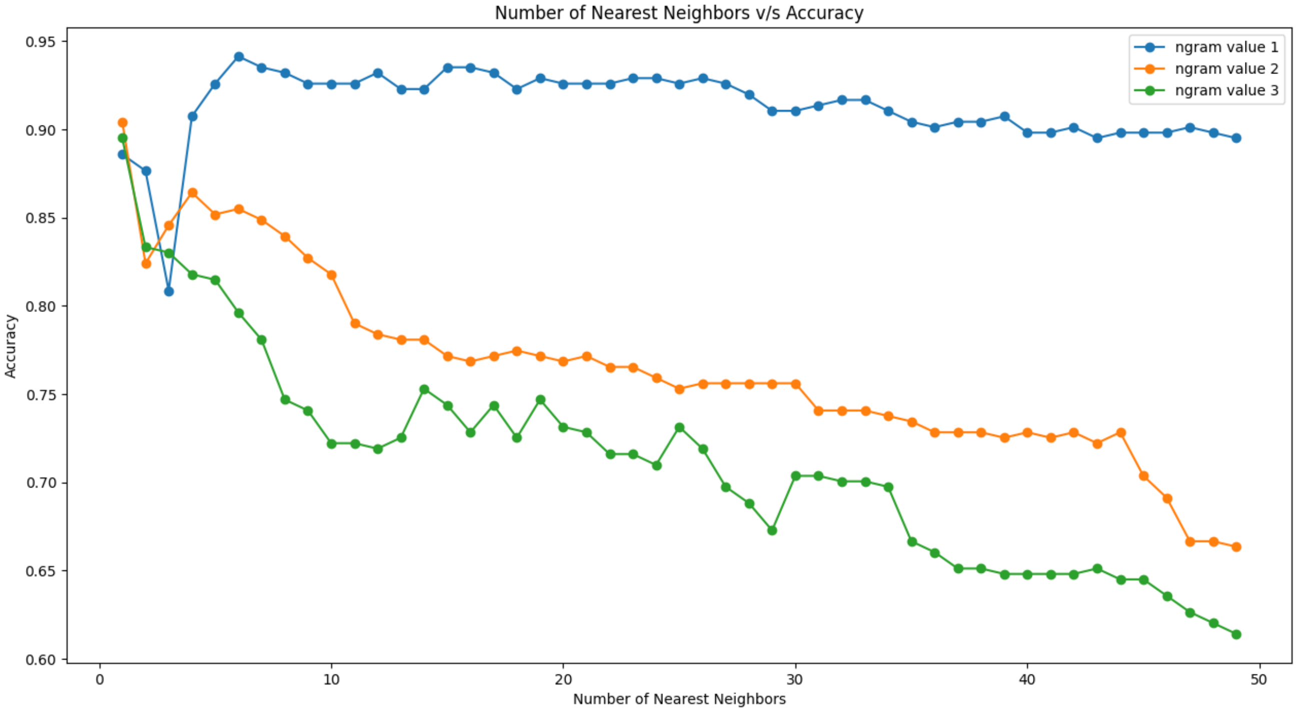
Maximum Depth v/s Accuracy for Decision Tree Classifier



Number of Classifiers v/s Accuracy for Random Forest Classifier



Number of Nearest Neighbors v/s Accuracy for KNN



Conclusion

The best accuracy was obtained when using the 1-gram data model. Both SVM with Sigmoid kernel and Random Forest Classifier gave similar accuracy of 98%.

Around 50 classifiers were found optimal after varying the number of classifiers. Though similar accuracy obtained for a greater number of classifiers also, when considering the convergence time, 50 gave the most optimal result.

The MLP converged after setting the maximum number of iterations to 150. Below that there weren’t enough iterations for the MLP to converge.

The Decision Tree Classifier’s maximum depth also was varied for checking the effect of pre-convergence of the tree. Increasing the depth further than 15-20 didn’t make much difference. Although some values gave slightly overfitted results.

For, K-Nearest Neighbors, the number of neighbours considered were varied. The accuracy decreased as the number of neighbours were increased from 5 onwards. Although 2-gram and 3-gram data models had a greater decline in accuracy, compared to the 1-gram model.

Data

MLP – Maximum Iterations v/s Accuracy

Maximum Iterations	**Accuracy**
1-gram:	
1	4.07E-01
101	9.51E-01
201	9.51E-01
301	9.51E-01
401	9.51E-01
501	9.51E-01
601	9.51E-01
701	9.51E-01
801	9.51E-01
901	9.51E-01
2-gram:	
1	5.56E-01
101	9.57E-01
201	9.51E-01
301	9.51E-01
401	9.51E-01
501	9.51E-01
601	9.51E-01
701	9.51E-01
801	9.51E-01
901	9.51E-01
3-gram:	
1	5.00E-01
101	9.41E-01
201	9.44E-01
301	9.44E-01
401	9.44E-01
501	9.44E-01
601	9.44E-01
701	9.44E-01
801	9.44E-01
901	9.44E-01

DTC – Maximum Depth v/s Accuracy

Maximum Depth	Accuracy
1-gram:	
1	4.10E-01
2	5.65E-01
3	6.88E-01
4	7.84E-01
5	7.93E-01
6	8.40E-01
7	8.67E-01
8	8.77E-01
9	8.92E-01
10	9.17E-01
11	9.26E-01
12	9.29E-01
13	9.35E-01
14	9.29E-01
15	9.26E-01
16	9.32E-01
17	9.32E-01
18	9.41E-01
19	9.32E-01
20	9.26E-01
21	9.44E-01
22	9.35E-01
23	9.29E-01
24	9.38E-01
25	9.35E-01
26	9.35E-01
27	9.41E-01
28	9.38E-01
29	9.35E-01
30	9.41E-01
31	9.35E-01
32	9.32E-01
33	9.35E-01
34	9.44E-01
35	9.44E-01
36	9.32E-01
37	9.41E-01
38	9.29E-01
39	9.51E-01
40	9.35E-01

41	9.38E-01
42	9.35E-01
43	9.44E-01
44	9.38E-01
45	9.48E-01
46	9.32E-01
47	9.41E-01
48	9.44E-01
49	9.44E-01
2-gram:	
1	4.14E-01
2	5.74E-01
3	6.42E-01
4	7.28E-01
5	7.90E-01
6	8.21E-01
7	8.58E-01
8	8.70E-01
9	8.92E-01
10	9.01E-01
11	9.14E-01
12	9.20E-01
13	8.92E-01
14	9.20E-01
15	9.20E-01
16	9.23E-01
17	9.29E-01
18	9.23E-01
19	9.29E-01
20	9.35E-01
21	9.29E-01
22	9.17E-01
23	9.32E-01
24	9.41E-01
25	9.41E-01
26	9.35E-01
27	8.77E-01
28	9.32E-01
29	9.44E-01
30	9.38E-01
31	9.32E-01
32	9.32E-01
33	9.35E-01
34	9.29E-01
35	9.32E-01

36	9.35E-01
37	9.41E-01
38	9.41E-01
39	9.35E-01
40	9.29E-01
41	9.32E-01
42	9.35E-01
43	9.32E-01
44	9.38E-01
45	9.29E-01
46	9.32E-01
47	9.35E-01
48	9.29E-01
49	9.32E-01
3-gram:	
1	4.14E-01
2	5.62E-01
3	6.33E-01
4	7.13E-01
5	7.59E-01
6	7.99E-01
7	8.12E-01
8	8.09E-01
9	8.21E-01
10	8.61E-01
11	8.70E-01
12	8.70E-01
13	8.36E-01
14	8.77E-01
15	8.80E-01
16	8.92E-01
17	9.04E-01
18	9.07E-01
19	9.07E-01
20	9.23E-01
21	9.20E-01
22	9.26E-01
23	9.23E-01
24	9.29E-01
25	8.98E-01
26	9.04E-01
27	9.10E-01
28	8.80E-01
29	9.10E-01
30	9.14E-01

31	8.83E-01
32	9.17E-01
33	9.17E-01
34	9.29E-01
35	9.23E-01
36	9.23E-01
37	9.23E-01
38	9.29E-01
39	9.23E-01
40	9.32E-01
41	9.29E-01
42	9.35E-01
43	9.38E-01
44	9.32E-01
45	9.32E-01
46	9.35E-01
47	9.32E-01
48	9.35E-01
49	9.32E-01

RFC – Number of Classifiers v/s Accuracy

Number of Classifiers	**Accuracy**
1-gram:	
1	8.89E-01
2	9.04E-01
3	9.10E-01
4	9.20E-01
5	9.48E-01
6	9.35E-01
7	9.44E-01
8	9.54E-01
9	9.38E-01
10	9.35E-01
11	9.35E-01
12	9.44E-01
13	9.51E-01
14	9.44E-01
15	9.48E-01
16	9.41E-01
17	9.57E-01
18	9.57E-01
19	9.54E-01
20	9.48E-01
21	9.51E-01

22	9.51E-01
23	9.54E-01
24	9.51E-01
25	9.41E-01
26	9.54E-01
27	9.60E-01
28	9.57E-01
29	9.44E-01
30	9.54E-01
31	9.54E-01
32	9.48E-01
33	9.54E-01
34	9.48E-01
35	9.51E-01
36	9.54E-01
37	9.48E-01
38	9.57E-01
39	9.44E-01
40	9.54E-01
41	9.51E-01
42	9.60E-01
43	9.54E-01
44	9.51E-01
45	9.51E-01
46	9.51E-01
47	9.48E-01
48	9.57E-01
49	9.54E-01
50	9.51E-01
51	9.57E-01
52	9.54E-01
53	9.44E-01
54	9.57E-01
55	9.48E-01
56	9.54E-01
57	9.48E-01
58	9.54E-01
59	9.48E-01
60	9.51E-01
61	9.51E-01
62	9.51E-01
63	9.60E-01
64	9.48E-01
65	9.48E-01
66	9.51E-01
67	9.54E-01

68	9.51E-01
69	9.57E-01
70	9.48E-01
71	9.51E-01
72	9.54E-01
73	9.48E-01
74	9.48E-01
75	9.60E-01
76	9.51E-01
77	9.57E-01
78	9.48E-01
79	9.48E-01
80	9.54E-01
81	9.51E-01
82	9.51E-01
83	9.51E-01
84	9.57E-01
85	9.51E-01
86	9.57E-01
87	9.57E-01
88	9.51E-01
89	9.51E-01
90	9.54E-01
91	9.51E-01
92	9.51E-01
93	9.51E-01
94	9.51E-01
95	9.54E-01
96	9.54E-01
97	9.54E-01
98	9.57E-01
99	9.51E-01
2-gram:	
1	8.70E-01
2	8.89E-01
3	9.10E-01
4	9.14E-01
5	9.32E-01
6	9.23E-01
7	9.32E-01
8	9.23E-01
9	9.32E-01
10	9.38E-01
11	9.38E-01

12	9.23E-01
13	9.23E-01
14	9.23E-01
15	9.29E-01
16	9.35E-01
17	9.38E-01
18	9.29E-01
19	9.44E-01
20	9.32E-01
21	9.38E-01
22	9.32E-01
23	9.29E-01
24	9.38E-01
25	9.38E-01
26	9.38E-01
27	9.35E-01
28	9.32E-01
29	9.35E-01
30	9.44E-01
31	9.35E-01
32	9.32E-01
33	9.38E-01
34	9.29E-01
35	9.32E-01
36	9.32E-01
37	9.32E-01
38	9.32E-01
39	9.35E-01
40	9.38E-01
41	9.41E-01
42	9.29E-01
43	9.29E-01
44	9.35E-01
45	9.35E-01
46	9.32E-01
47	9.35E-01
48	9.41E-01
49	9.38E-01
50	9.35E-01
51	9.38E-01
52	9.32E-01
53	9.35E-01
54	9.32E-01
55	9.35E-01
56	9.38E-01
57	9.35E-01

58	9.29E-01
59	9.35E-01
60	9.35E-01
61	9.32E-01
62	9.38E-01
63	9.29E-01
64	9.35E-01
65	9.32E-01
66	9.32E-01
67	9.32E-01
68	9.35E-01
69	9.35E-01
70	9.38E-01
71	9.32E-01
72	9.29E-01
73	9.32E-01
74	9.38E-01
75	9.32E-01
76	9.41E-01
77	9.35E-01
78	9.35E-01
79	9.35E-01
80	9.32E-01
81	9.26E-01
82	9.35E-01
83	9.32E-01
84	9.38E-01
85	9.38E-01
86	9.29E-01
87	9.41E-01
88	9.35E-01
89	9.32E-01
90	9.35E-01
91	9.32E-01
92	9.38E-01
93	9.35E-01
94	9.29E-01
95	9.32E-01
96	9.29E-01
97	9.32E-01
98	9.35E-01
99	9.38E-01
3-gram:	
1	8.92E-01

2	8.80E-01
3	9.44E-01
4	9.07E-01
5	9.04E-01
6	9.10E-01
7	9.26E-01
8	9.20E-01
9	9.17E-01
10	9.32E-01
11	8.95E-01
12	9.20E-01
13	9.17E-01
14	9.10E-01
15	9.26E-01
16	9.26E-01
17	9.17E-01
18	9.23E-01
19	9.23E-01
20	9.29E-01
21	9.29E-01
22	9.26E-01
23	9.26E-01
24	9.29E-01
25	9.26E-01
26	9.26E-01
27	9.20E-01
28	9.17E-01
29	9.29E-01
30	9.23E-01
31	9.26E-01
32	9.26E-01
33	9.26E-01
34	9.26E-01
35	9.23E-01
36	9.23E-01
37	9.23E-01
38	9.26E-01
39	9.26E-01
40	9.26E-01
41	9.23E-01
42	9.20E-01
43	9.26E-01
44	9.20E-01
45	9.26E-01
46	9.26E-01
47	9.26E-01

48	9.23E-01
49	9.23E-01
50	9.26E-01
51	9.29E-01
52	9.26E-01
53	9.26E-01
54	9.20E-01
55	9.26E-01
56	9.26E-01
57	9.23E-01
58	9.29E-01
59	9.14E-01
60	9.23E-01
61	9.26E-01
62	9.26E-01
63	9.29E-01
64	9.26E-01
65	9.26E-01
66	9.26E-01
67	9.26E-01
68	9.26E-01
69	9.26E-01
70	9.26E-01
71	9.26E-01
72	9.26E-01
73	9.26E-01
74	9.20E-01
75	9.26E-01
76	9.26E-01
77	9.26E-01
78	9.26E-01
79	9.26E-01
80	9.26E-01
81	9.23E-01
82	9.26E-01
83	9.29E-01
84	9.26E-01
85	9.26E-01
86	9.26E-01
87	9.26E-01
88	9.26E-01
89	9.26E-01
90	9.26E-01
91	9.26E-01
92	9.26E-01
93	9.29E-01

94	9.26E-01
95	9.26E-01
96	9.26E-01
97	9.26E-01
98	9.26E-01
99	9.26E-01

KNN – Number of Nearest Neighbors v/s Accuracy

Number of Nearest Neighbors	**Accuracy**
1 gram:	
1	8.86E-01
2	8.77E-01
3	8.09E-01
4	9.07E-01
5	9.26E-01
6	9.41E-01
7	9.35E-01
8	9.32E-01
9	9.26E-01
10	9.26E-01
11	9.26E-01
12	9.32E-01
13	9.23E-01
14	9.23E-01
15	9.35E-01
16	9.35E-01
17	9.32E-01
18	9.23E-01
19	9.29E-01
20	9.26E-01
21	9.26E-01
22	9.26E-01
23	9.29E-01
24	9.29E-01
25	9.26E-01
26	9.29E-01
27	9.26E-01
28	9.20E-01
29	9.10E-01
30	9.10E-01
31	9.14E-01
32	9.17E-01
33	9.17E-01
34	9.10E-01

35	9.04E-01
36	9.01E-01
37	9.04E-01
38	9.04E-01
39	9.07E-01
40	8.98E-01
41	8.98E-01
42	9.01E-01
43	8.95E-01
44	8.98E-01
45	8.98E-01
46	8.98E-01
47	9.01E-01
48	8.98E-01
49	8.95E-01
2-gram:	
1	9.04E-01
2	8.24E-01
3	8.46E-01
4	8.64E-01
5	8.52E-01
6	8.55E-01
7	8.49E-01
8	8.40E-01
9	8.27E-01
10	8.18E-01
11	7.90E-01
12	7.84E-01
13	7.81E-01
14	7.81E-01
15	7.72E-01
16	7.69E-01
17	7.72E-01
18	7.75E-01
19	7.72E-01
20	7.69E-01
21	7.72E-01
22	7.65E-01
23	7.65E-01
24	7.59E-01
25	7.53E-01
26	7.56E-01
27	7.56E-01
28	7.56E-01
29	7.56E-01

30	7.56E-01
31	7.41E-01
32	7.41E-01
33	7.41E-01
34	7.38E-01
35	7.35E-01
36	7.28E-01
37	7.28E-01
38	7.28E-01
39	7.25E-01
40	7.28E-01
41	7.25E-01
42	7.28E-01
43	7.22E-01
44	7.28E-01
45	7.04E-01
46	6.91E-01
47	6.67E-01
48	6.67E-01
49	6.64E-01
3-gram:	
1	8.95E-01
2	8.33E-01
3	8.30E-01
4	8.18E-01
5	8.15E-01
6	7.96E-01
7	7.81E-01
8	7.47E-01
9	7.41E-01
10	7.22E-01
11	7.22E-01
12	7.19E-01
13	7.25E-01
14	7.53E-01
15	7.44E-01
16	7.28E-01
17	7.44E-01
18	7.25E-01
19	7.47E-01
20	7.31E-01
21	7.28E-01
22	7.16E-01
23	7.16E-01
24	7.10E-01

25	7.31E-01
26	7.19E-01
27	6.98E-01
28	6.88E-01
29	6.73E-01
30	7.04E-01
31	7.04E-01
32	7.01E-01
33	7.01E-01
34	6.98E-01
35	6.67E-01
36	6.60E-01
37	6.51E-01
38	6.51E-01
39	6.48E-01
40	6.48E-01
41	6.48E-01
42	6.48E-01
43	6.51E-01
44	6.45E-01
45	6.45E-01
46	6.36E-01
47	6.27E-01
48	6.20E-01
49	6.14E-01