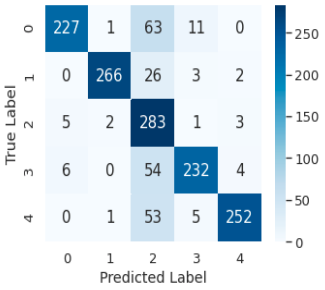
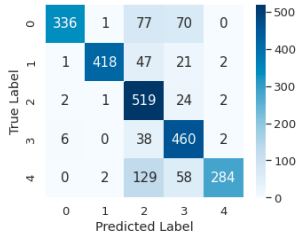
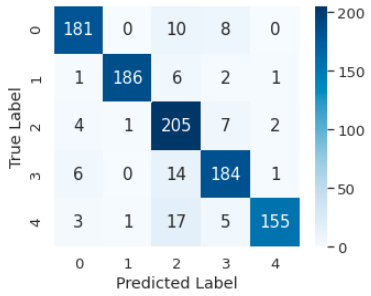
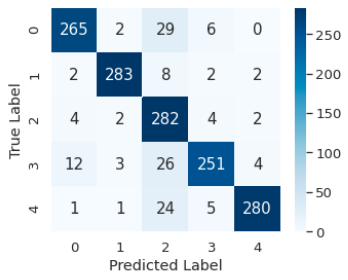
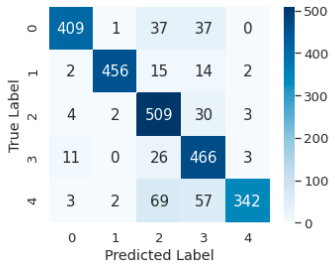
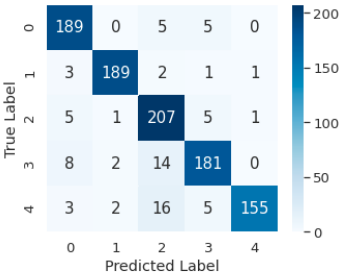
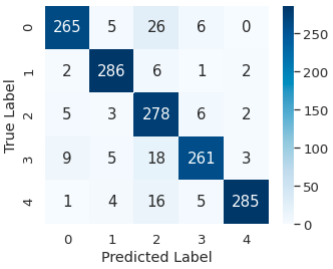
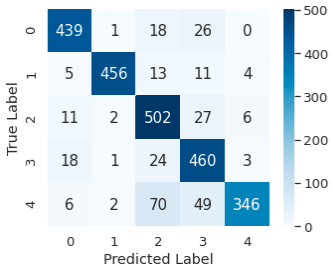
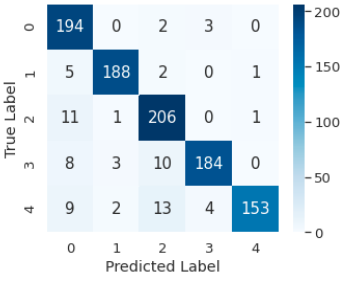
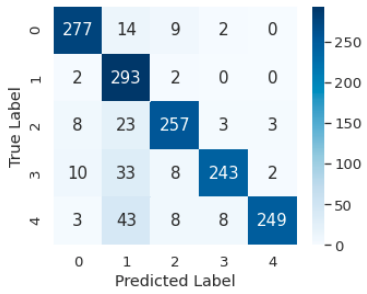
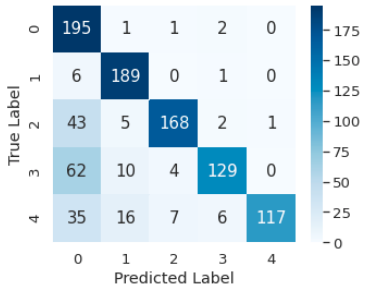
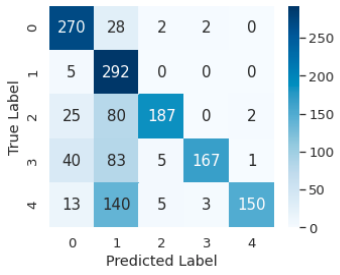
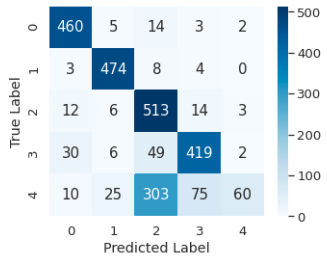
Analysis Doc

Q1:

Number of features (top-k)	Accuracy	Confusion matrix	Split
300	85.3	 Class: ['comp.graphics'] Accuracy: 155/162 Class: ['rec.sport.hockey'] Accuracy: 172/175 Class: ['sci.med'] Accuracy: 212/309 Class: ['sci.space'] Accuracy: 186/224 Class: ['talk.politics.misc'] Accuracy: 128/130	80-20
300	84	 Class: ['comp.graphics'] Accuracy: 227/238 Class: ['rec.sport.hockey'] Accuracy: 266/270 Class: ['sci.med'] Accuracy: 283/479 Class: ['sci.space'] Accuracy: 232/252	70-30

		Class: ['talk.politics.misc'] Accuracy: 252/261	
300	80.68	 Class: ['comp.graphics'] Accuracy: 336/345 Class: ['rec.sport.hockey'] Accuracy: 418/422 Class: ['sci.med'] Accuracy: 519/810 Class: ['sci.space'] Accuracy: 460/633 Class: ['talk.politics.misc'] Accuracy: 284/290	50-50
700	91.1		80-20
700	90.7		70-30

700	87.28		50-50
1000	92.1		80-20
1000	91.66		70-30
1000	88.12		50-50
1500	92.5		80-20

1500	87.93		70-30
1500	72.88		50-50
1800	79.8		80-20
1800	71.06		70-30
1800	77.06		50-50

Analysis:
We observe that at top-300 features for each class we get 85.3 as highest accuracy for 80-20 split. For top-1000 features we reach 92.1% accuracy on 80-20 split and also the accuracies are fairly good for 70-30 and 50-50 splits as seen from table. However the performance deteriorates as we move to top 1800 features. This indicates top-1000 features for each class is optimal for the given data.

Q2 - Part - 1)

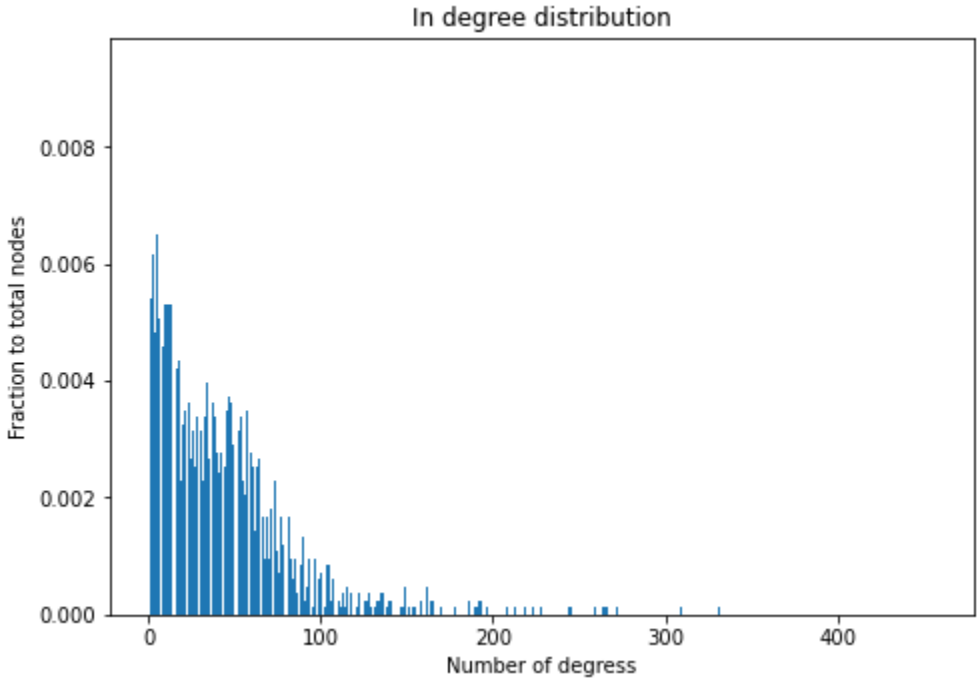
Following are the characteristics (output) of the dataset:

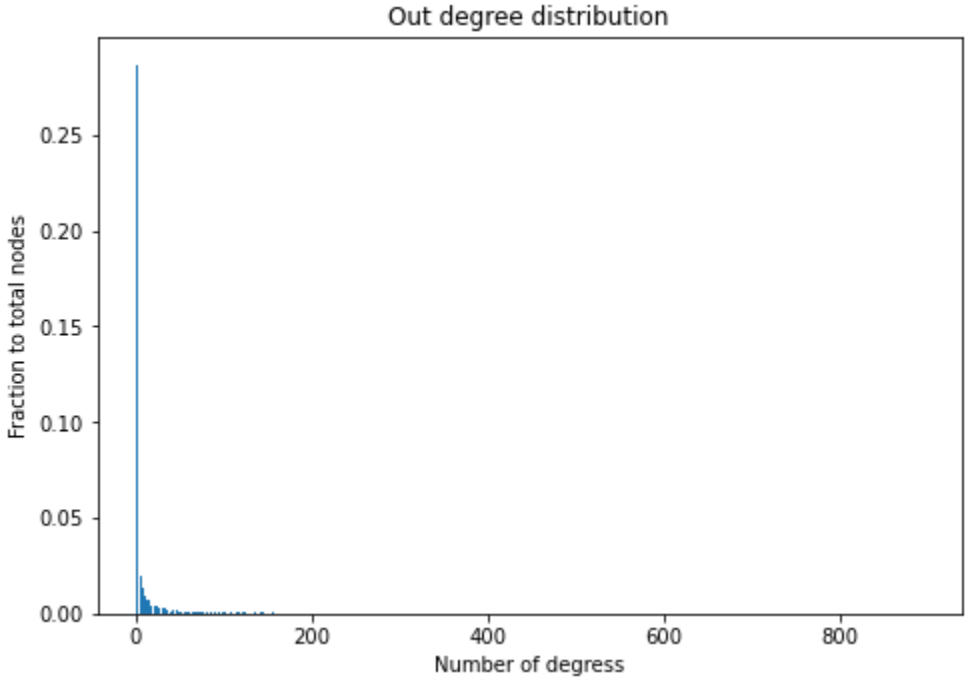
No of nodes: 7115
No of edges: 103689
Average in-degree: 16.97037643207856
Average out-degree: 43.54850902981941
Density of network: 0.15064088296719116

Clustering coefficient (sample set of values):
0: 0
1: 0
2: 0
3: 1.06533892382949
4: 1.1724137931034482
5: 1.2371541501976284
6: 1.01029391846133
7: 1.1358695652173914
8: 1.0115240904621436
9: 1.0666666666666667
10: 1.062871287128713
11: 1.027469318309614
12: 1.0668269230769232
13: 1.2181818181818183
14: 1.03959012575687
15: 1.013352323304255
16: 1.1904761904761905
17: 1.1106060606060606
18: 1.2196969696969697
19: 1.0885245901639344
20: 1.038439422689897
21: 1.043859649122807
22: 2.0
23: 1.044407894736842
24: 1.0451498339847853
25: 1.1119850187265918
26: 1.0712206047032475

27: 1.0479518540089303
28: 0.9422109001080747
29: 0.9904656930283825
30: 1.1177248677248677
31: 1.1238095238095238
32: 1.0816993464052287
33: 1.0907441016333939
34: 1.0933333333333333
35: 0.9805582922824302
36: 1.0253315649867374
37: 1.1210526315789473
38: 1.1825396825396826
39: 1.011082693947144
40: 1.190909090909091
41: 1.106896551724138
42: 1.0481358529111338
43: 1.4166666666666667
44: 1.0586680761099365
45: 1.3035714285714286
46: 1.0669199298655756
47: 1.0568256967406708
48: 1.161904761904762
49: 1.0743243243243243
50: 1.1198067632850242
51: 1.1304347826086956
52: 0
53: 1.0897435897435896
54: 1.1063829787234043
55: 0.9810921902524956
56: 1.0279271407447237
57: 1.1794871794871795
58: 1.1159317211948792
59: 1.042438801189659
60: 1.2361111111111112
61: 1.25
62: 1.1433823529411764
63: 0
64: 1.3666666666666667
65: 2.5
66: 1.0845665961945032
67: 1.1580882352941178
68: 1.047193098871931
69: 0
70: 0

71: 1.0538566089024806
72: 0.9983334325337777
73: 1.2564102564102564
74: 1.1010230179028133





In degree Centrality - (Sample few nodes)

In-Degree Centrality: {4: 0.06783369803063458, 7: 0.0437636761487965, 9: 0.0962800875273523, 11: 0.03282275711159737, 16: 0.7899343544857768, 20: 0.04814004376367615, 24: 0.04814004376367615, 29: 0.26695842450765866, 30: 0.08971553610503283, 31: 0.05032822757111598, 33: 0.030634573304157548, 34: 0.07439824945295405, 35: 0.0437636761487965, 36: 0.24070021881838075, 37: 0.07439824945295405, 39: 0.030634573304157548, 40: 0.061269146608315096, 50: 0.04814004376367615, 51: 0.0700218818380744, 55: 0.087527352297593, 56: 0.14660831509846828, 57: 0.32603938730853393, 62: 0.0437636761487965, 65: 0.019693654266958426, 73: 0.33698030634573306, 76: 0.15098468271334792, 81: 0.09409190371991247, 87: 0.2363238512035011, 90: 0.0350109409190372, 94: 0.037199124726477024, 95: 0.05908096280087528, 96: 0.08533916849015317, 106: 0.045951859956236324 …

Out degree Centrality - (Sample few nodes)

Out-Degree Centrality: {4: 0.025755879059350503, 5: 0.032474804031354984, 6: 0.025755879059350503, 7: 0.3381858902575588, 8: 0.026875699888017916, 9: 0.2038073908174692, 10: 0.09070548712206047, 11: 0.09630459126539753, 12: 0.832026875699888, 13: 0.07278835386338185, 14: 0.012318029115341545, 15: 0.1276595744680851, 16: 0.055991041433370664, 17: 0.0167973124300112, 18: 0.05039193729003359, 19: 0.013437849944008958, 20: 0.043673012318029114, 21: 0.20940649496080627, 22: 0.0851063829787234, 23: 0.0022396416573348264, 24: 0.08286674132138858, 25: 0.3460246360582307, 26: 0.10078387458006718, 27: 0.10638297872340426, 28: 0.11422172452407615, 29: 0.14893617021276595, 30: 0.167973124300112, 31: 0.005599104143337066, 32: 0.0167973124300112, 33: 0.004479283314669653, 34: 0.026875699888017916, 35: 0.005599104143337066, 36: 0.07278835386338185, 37: 0.2541993281075028, 38: 0.022396416573348264, 39: 0.015677491601343786, 40: 0.04591265397536394, 41: 0.012318029115341545, 42: 0.0335946248600224, 43: 0.09966405375139978, 44: 0.004479283314669653, 45: 0.04927211646136618, 46: 0.008958566629339306, 47: 0.06606942889137737, 48: 0.16349384098544234, 49: 0.023516237402015677, 50: 0.0167973124300112, 51: 0.015677491601343786, 52: 0.026875699888017916, 53: 0.0011198208286674132, 54: 0.043673012318029114, 55: 0.007838745800671893, 56: 0.07166853303471445, 57: 0.026875699888017916, 58: 0.014557670772676373, 59: 0.0425531914893617, 60: 0.10526315789473684, 61: 0.010078387458006719, 63: 0.019036954087346025, 64: 0.0011198208286674132, 65: 0.0011198208286674132, 66: 0.0022396416573348264, 67: 0.04927211646136618, 68: 0.019036954087346025, 69: 0.3079507278835386, 72: 0.12206047032474804, 73: 0.18365061590145576, 74: 0.014557670772676373, 75: 0.07726763717805152, 76: 0.03471444568868981, 77: 0.0022396416573348264, 78: 0.027995520716685332, 79: 0.04591265397536394, 80: 0.0593505039193729, 81: 0.06606942889137737, 82: 0.020156774916013438, 83: 0.0011198208286674132, 84: 0.008958566629339306, 85: 0.0011198208286674132, 86: 0.0011198208286674132, 87:

0.1366181410974244, 88: 0.21836506159014557, 89: 0.0011198208286674132, 90: 0.032474804031354984, 91: 0.08062709966405375, 92: 0.0022396416573348264, 93: 0.005599104143337066, 94: 0.0011198208286674132, 95: 0.08958566629339305, 96: 0.0022396416573348264, 97: 0.0022396416573348264, 99: 0.008958566629339306, 100: 0.026875699888017916, 101: 0.005599104143337066, 102: 0.01791713325867861, 103: 0.005599104143337066, 104: 0.0761478163493841, 105: 0.0335946248600224, 106: 0.0033594624860022394 …

Question 2 Part -2

We display only the top 5 outputs here and the scores of all nodes

Pagerank:
top 5 nodes with pagerank scores are: [(0.003992602203050151, 4037), (0.003616213680639039, 15), (0.003116042657586646, 2625), (0.0026539389518222623, 6634), (0.002595552667734392, 7553)]

HITS:
Hub scores:
top 5 nodes according to hub scores are: [(0.007940491702773854, 2565), (0.0075743344856463975, 766), (0.00644024862307382, 2688), (0.006416869643642841, 457), (0.006010568182523095, 1166)]

Authorities:
top 5 nodes according to authorities scores are: [(0.0025801471013370126, 2398), (0.0025732399689523578, 4037), (0.0023284155432415415, 3352), (0.0023037316962506937, 1549), (0.0022558759994645286, 762)]

Analysis:
We observe that there is only one common highly ranked node “4037” between Pagerank and Authorities based ranking. Authorities are aggregated based on incoming links and hub scores are compute dusing outgoing links. The ranking is very different between Pagerank and HITS because in HITS there is a circular definition where hubs reinforce authorities and authorities reinforce hubs. The values could go unbounded and the nodes with large number of outlinks influence the scores. But in pagerank the score of a node is equally divided between all the outgoing links and computes the scores of nodes using random walk model. In pagerank a node with sufficient number of inlinks has a higher score.