Report: Project 2

Utkarsh Kajaria

The extra preprocessing steps done are:

a) Removing all the words/n-grams that are of length that are equal to 1

b) Removing the stop words.

Statistics:

Representation	Number of Objects	Non-Zero Dimensions	Total number of Features
Bag of Words	6521	701550	100758
3-gram	6521	2891332	41908
5-gram	6517	5169836	577260
7-gram	6517	6234672	1834321

Evaluation of Quality of the clustering algorithm and runtime efficiency:-

Representation	Clusters	Entropy	Purity	Time for 20 iterations
BagofWords	20	-2.712	0.4099	1480
BagofWords	40	-2.622	0.412	151
BagofWords	60	-2.596	0.42	223
3-gram	20	-3.177	0.287	228
3-gram	40	-2.983	0.32	296
3-gram	60	-2.78	0.34	538
5-gram	20	-2.818	0.356	600
5-gram	40	-2.623	0.38	1252
5-gram	60	-2.474	0.408	1620
7-gram	20	-3.014	0.33	1100
7-gram	40	-2.824	0.36	1665
7-gram	60	-2.766	0.39	2292

From the above analysis, we can see that purity increases with increase in the number of clusters for the same input file. We also see that the purity is minimum for the 3-gram representation among all the representation and also has the least number of distinct features.