REPORT FOR SPHERICAL K MEANS CLUSTERING

Data Preprocessing

Data Preprocessing was done as per required to produce the dimensions and representations. No additional steps apart from the ones mentioned in the problem statement was followed. The following results were obtained.

	Dimensionality of feature space	**No. of non zeroes**
Bag of Words	6788	485851
3 Grams	5531	2497320
5 Grams	52922	3666414
7 Grams	127748	3601763

The preprocessing runs in **76.39 seconds**

The Objective Function

The objective function that was to be **maximised** was cosine similarity to the vectors. This means that we need to find the value of the following function

$\Sigma_{k \in Clusters}$ cosine_similarity (Ck, articles of that cluster)

Clustering

REPRESENTATION	**Objective fn,**	**Entropy**	**Purity**	**Exec. Time(s)**
Bag , 20 Clusters	6036.165	0.567	0.731	206
Bag, 40 Clusters	6155.221	0.603	0.755	301
Bag, 60 Clusters	6237.61	0.322	0.771	504
3 Gram, 20 Clusters	5958.342	0.600	0.726	2311
3 Gram, 40 Clusters	6097.47	0.391	0.772	3101
3 Gram, 60 Clusters	6212.412	0.388	0.798	4041

5 Gram, 20 Clusters	4378.65	0.601	0.743	4581
5 Gram, 40 Clusters	4582.35	0.354	0.783	6944
5 Gram, 60 Clusters	4727.983	0.314	0.811	7199
7 Gram, 20 Clusters	3279.516	0.544	0.746	3318
7 Gram, 40 Clusters	3528.574	0.371	0.781	4693
7 Gram, 60 Clusters	3733.828	0.301	0.804	7660

Observation

This means that as the dimensionality of our feature space increases, the value of the objective function will fall. However, the purities and entropies will go up with both no. of clusters and with more non-zero feature representations. As observed, no of non zero features is lower for 7 grams than 5 grams and hence the observed anomaly.

Result

Thus the spherical k means algorithm was implemented for the given dataset with resulting clusters of purities approximately 0.80.