CS6370 Programming Assignment - IV

Sabarinath N P (CS10B020)

Introduction

The objective of this assignment is to perform LDA on the given document collection and use it to cluster the documents. Gibbs sampling method is used as the subroutine to perform LDA. The purity of this topic model is then compared against the base model.

Dataset Description

- The dataset contains 2761 documents each belonging one of 5 topics
- Size of vocabulary is 46885

Theory

In each round the topic for each word is chosen using Gibbs sampling assuming topic is assigned to all other words in a document. The probability distribution is given by

$$P(z = k|.) = (n_{d,k} + \alpha_k) * \frac{n_{k,w} + \beta_w}{n_k + \beta * |V|}$$

where,
$n_{d,k}$ is number of words are assigned with topic k in document d
$n_{k,w}$ is number of times word w is assgined topic k
n_k is the total number of words assgined with topic k
$|V|$ is the size of vocabulary

The parameters are taken according to the given LDA reference. $\alpha = \frac{50}{|K|}$ and $\beta = 0.01$ In the second part of the assignment the goodness of clustering is measure in terms of purity which is given as follows,

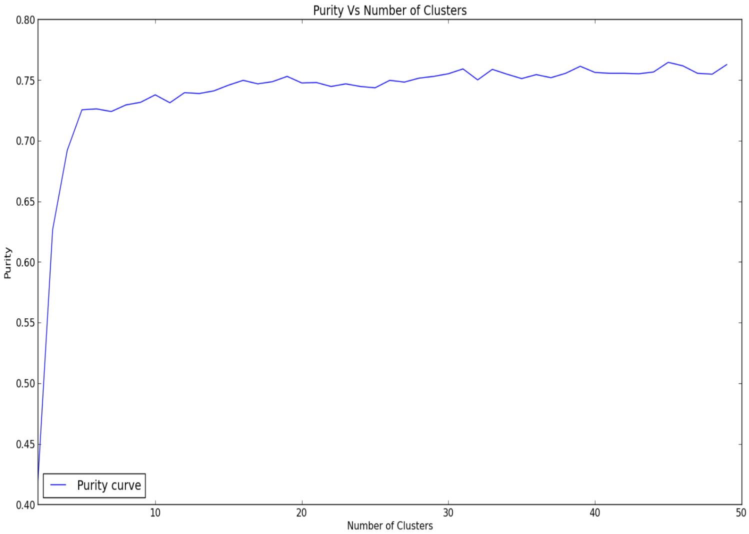
$$purity(\Omega, \mathcal{C}) = \frac{1}{N} \sum_k max_j |\omega_k \cap c_j|$$

Results

When number of topics is fixed to 5 and the number of clusters in K-Means is varied from 2 to 49, the cluster purity gradually increases after an inital burst, from 41% to 76% (when gibbs smapling run for 1000 iterations). This is mainly attributed by the fact that as the nuber of clusters increases the cluster size decreases and hence the number of documents assgined with wrong topic decreases faster than number of correctly assigned documents. The below table displays the first 10 different number of clusters

Number of Clusters	Purity
2	0.419
3	0.626
4	0.692
5	0.725
6	0.726
7	0.723
8	0.729
9	0.731
10	0.737
11	0.739

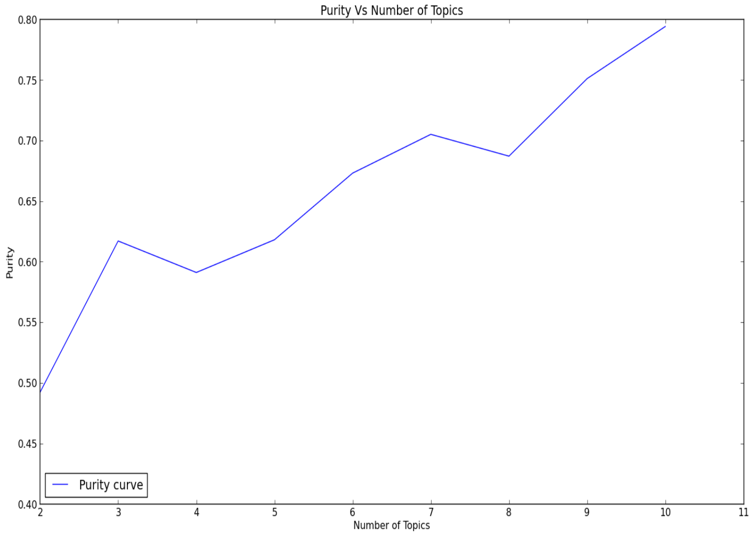
The corresponding graph,



When number of clusters is fixed to 5 and the number of topics is varied from 2 to 10, the cluster purity increases to 79% from 49% though there are osillations in the purity value. The below table shows the recorded purity values.

Number of Topics	Purity
2	0.492
3	0.617
4	0.591
5	0.618
6	0.673
7	0.705
8	0.687
9	0.751
10	0.794

The corresponding graph,



For instance,

- Cluster purity when number of topics is 5, number of clusters is 5 – 72%
- Cluster purity when number of topics is 4, number of clusters is 5 – 59.1%

The top 5 words belonging to each topic. Here topic need not necessarily mean a topic with name. We just consider each topic as a distinct number. And we are just interested in finding the words that belong to same topic as finding the actual topic of a given word is a different problem which is out of scope for this assignment. Below table contains the first 5 words of each topic.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
nt	x	nt	organization	x
re	nt	one	subject	window
game	organization	people	lines	use
writes	ones	would	university	file
year	would	like	nntpostinghost	program

It can be seen that the words other than the repeating stop words, belonging to same topic have the same meaning and add to the semantics of the topic.

When this topic model is compared against the base model where KMeans is run on the binary vector representation of each document, this model performs well in terms of purity values. The base model achieved purity value of 24.3% whereas our topic model achieved 70% on an average.

Conclusion

As discussed in this report, LDA serves as a powerful tool in grouping the words that belong the same topic and hence assign topics to the documents which any requirement of the actual semantics.