Assignment 4 :

Task1 : We have performed LDA using Gibbs sampling and foe each topic we have written the 5 most frequent words in the file tokpicwords

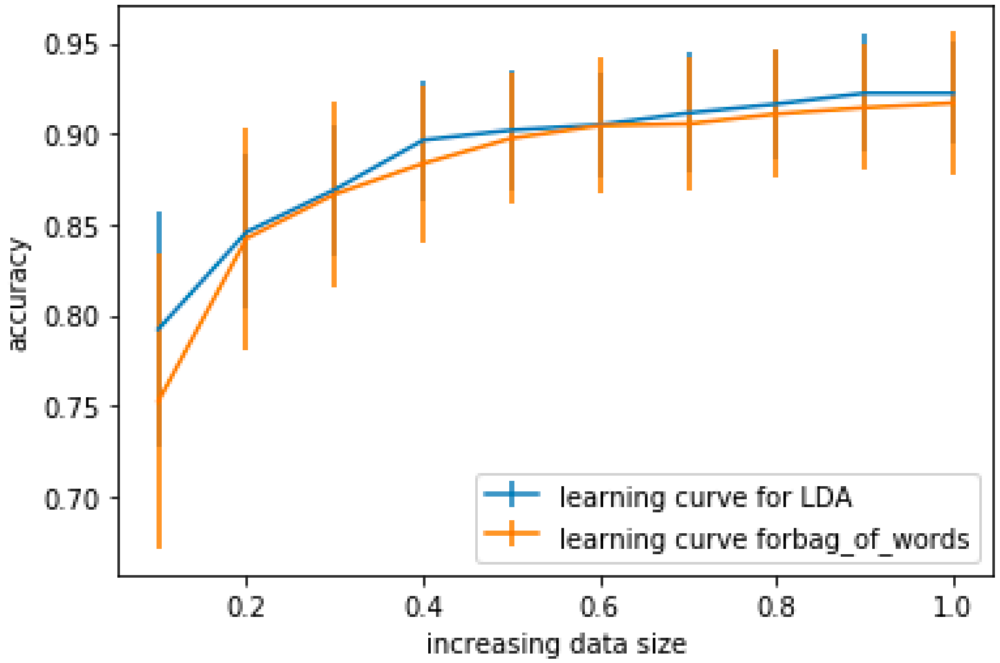
Result for the output of task1 :

	A	B
		topicwords
1	**0**	**['system', 'part', 'mars', 'earth', 'orbit']**
2	**1**	['car', 'ford', 'probe', 'mustang', 'drive']
3	**2**	['edu', 'gif', 'uci', 'ics', 'incoming']
4	**3**	['engine', 'feel', 'toyota', 'ford', 'seat']
5	**4**	['clutch', 'shifter', 'manual', 'sho', 'cars']
6	**5**	['writes', 'article', 'edu', 'apr', 'oort']
7	**6**	['sky', 'light', 'rights', 'news', 'moon']
8	**7**	['make', 'find', 'good', 'even', 'two']
9	**8**	['oil', 'lights', 'service', 'change', 'come']
10	**9**	['time', 'used', 'never', 'another', 'come']
11	**10**	['edu', 'writes', 'article', 'apr', 'engines']
12	**11**	['don', 'insurance', 'want', 'geico', 'cost']
13	**12**	['henry', 'edu', 'writes', 'toronto', 'spencer']
14	**13**	['edu', 'information', 'internet', 'resources', 'writes']
15	**14**	['bill', 'moon', 'george', 'eliot', 'howell']
16	**15**	['etc', 'day', 'lot', 'light', 'air']
17	**16**	['earth', 'large', 'temperature', 'things', 'mars']
18	**17**	['space', 'nasa', 'long', 'program', 'gov']
19	**18**	['station', 'shuttle', 'option', 'launch', 'two']
20	**19**	['hst', 'mission', 'pat', 'access', 'net']

Here, from results in topicwords, we are getting similar type of words in similar topic :
For eg : sky , moon , light
->car , ford , drive , mustang
-> clutch , shifter , manual
-> earth, temperature , mars.
Hence from the above examples we can see that all these words are associated with each other in some similar way.

Task2 :
Here , in this task we have to take to methods, LDA approach using Gibbs sampling and bag of words approach and compare their logistic regression accuracy on the test data for increasing data size.

Result of graph :



From the above graph, we can say that the LDA approach has better accuracy than the bag of words approach.
Here, the accuracy for LDA approach is around 92-93%, Whereas the accuracy for the bag of words is around 90%.