MACHINE LEARNING PROGRAMMING PROJECT 4:

This Project was completed in RStudio.

We were given 2 datasets for this project:

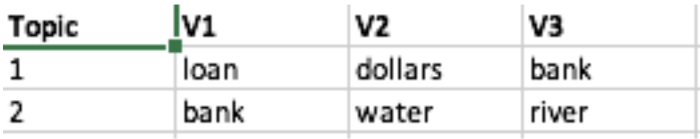
1. Artificial Data
2. Twenty Newsgroup Data

We made our LDA algorithm which included the Gibbs Sampling procedure. This algorithm included updating the Document Topic matrix and word topic count matrix 2by resampling the topics on each iteration using the new probability calculated at each step.

Our final word topic count matrix gave us the most frequent words for each topic. This matrix is returned by our main function into the working directory/pp4data/”artificial” OR “20newsgroups” folder as per the requirement of the question.

1. Artificial Data: These are the 3 most frequent words in topic 1 & 2.

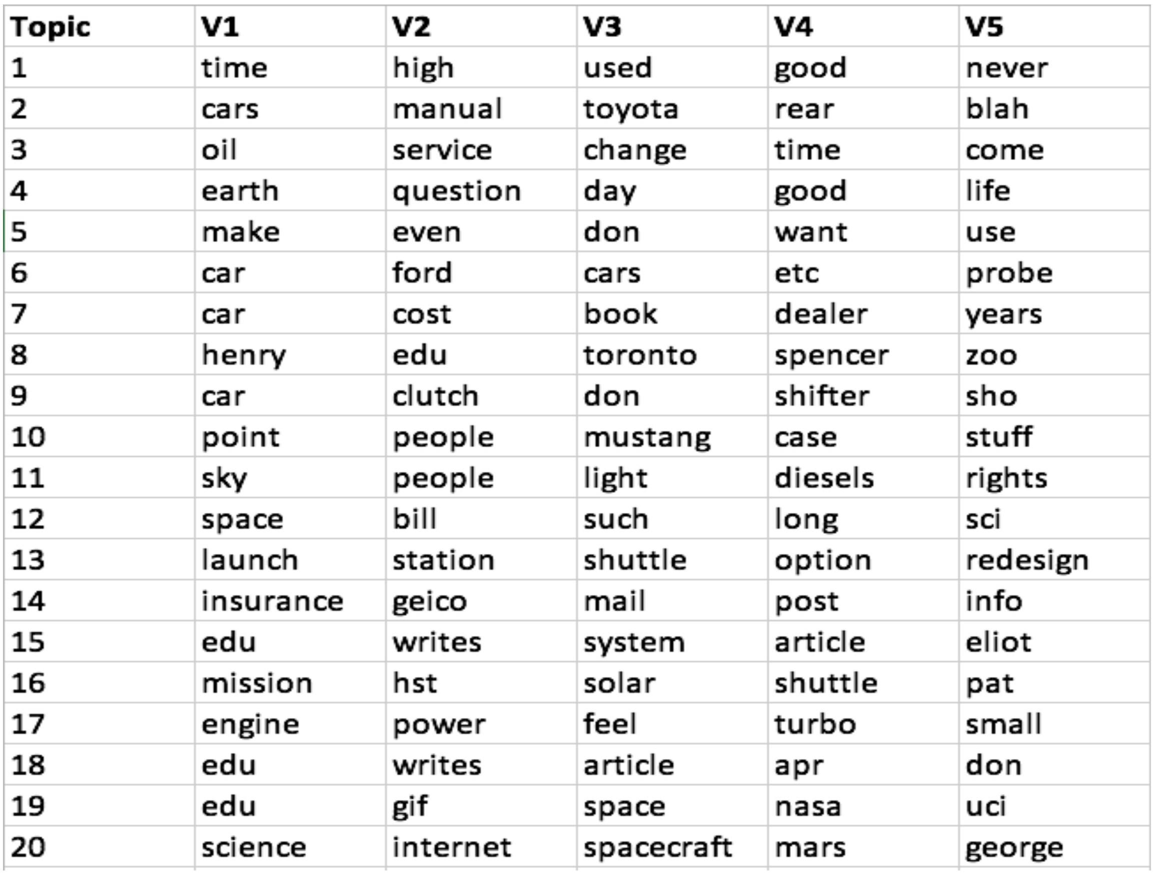
   About the topics making sense, both the topics make sense in the artificial data. The first topic relates to some banking terms and the 2nd topic relates to nature and river.



Topic	V1	V2	V3
1	loan	dollars	bank
2	bank	water	river

2. 20Newsgroup Data: These are the 5 most frequent words in topics 1 to 20 in this dataset.

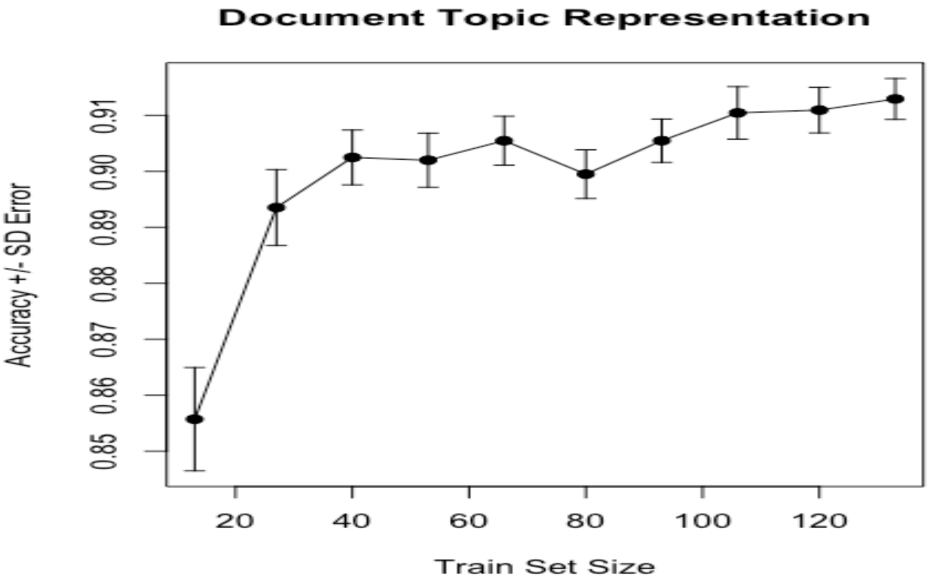
Here too most of the topics make sense. For eg: Topics 2,6,9,17 are topics about cars with most frequent words in it being cars, Toyota, ford, clutch, shifter, manual, rear, engine turbo, power. Topic 7 is about car dealers with words like car, dealer cost, book, years. Topics 16,19,20 are about science articles with words like solar, shuttle, space, Nasa, Science, internet, mars, spacecraft. So, we can see that almost all the topics have an inherent theme about them which is evident from the most frequent words in each topic.



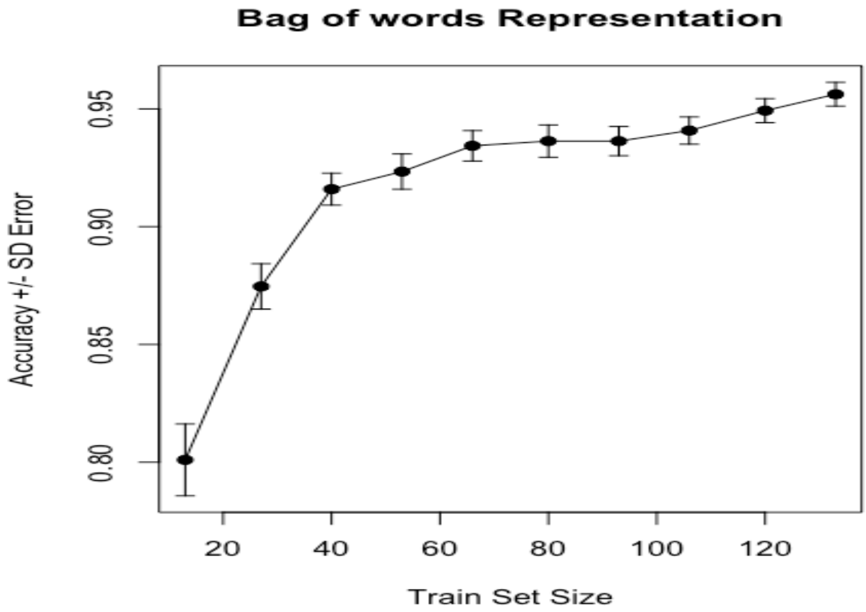
Topic	V1	V2	V3	V4	V5
1	time	high	used	good	never
2	cars	manual	toyota	rear	blah
3	oil	service	change	time	come
4	earth	question	day	good	life
5	make	even	don	want	use
6	car	ford	cars	etc	probe
7	car	cost	book	dealer	years
8	henry	edu	toronto	spencer	zoo
9	car	clutch	don	shifter	sho
10	point	people	mustang	case	stuff
11	sky	people	light	diesels	rights
12	space	bill	such	long	sci
13	launch	station	shuttle	option	redesign
14	insurance	geico	mail	post	info
15	edu	writes	system	article	eliot
16	mission	hst	solar	shuttle	pat
17	engine	power	feel	turbo	small
18	edu	writes	article	apr	don
19	edu	gif	space	nasa	uci
20	science	internet	spacecraft	mars	george

Now, as per task 2 of Classification, we had to predict the labels of index.csv file based on the final document topic matrix derived from our LDA Algorithm after Gibbs sampling above. We had to calculate the Document Topic Distribution of our document topic matrix and run our logistic regression function which we made in programming assignment 3 using this document topic distribution and then predict the labels of index.csv file.

Here is my accuracy graph based on 30 iterations of our logistic regression function on the document topic representation for various train set sizes.



Then we had to run our logistic regression function on the bag of words matrix which is basically a count of unique words for each document of the 20newsgroups dataset and then predict the labels of the index.csv file using this bag of words model.
Here is my accuracy graph based on 30 iterations of our logistic regression function on the bag of words model for various train set sizes.



Discussion on the results of these 2 graphs:

We see that the accuracy of our LDA Document topic representation (0.91) is almost the same as the bag of words representation (0.95) in predicting the labels of each of the 200 documents of the 20newsgroups dataset. The bag of words model uses 405 variables(unique words) count for each document to predict the labels of each document whereas LDA uses dimensionality reduction and only uses 20 topics i.e. 20 variables to predict the labels of each of the 200 documents. So, we can say that the dimensionality reduction in case of LDA works just as well the bag of words model in predicting the labels with negligible run time in running the logistic regression function due to less variables present in LDA(20 variables). These 20 variables or the TOPICS FOR EACH DOCUMENT work pretty well in representing and explaining each document and that is why we were able to get similar accuracy for predicting the labels of each document using both Document Topic Representation and the Bag of words Model.