PROGRAMMING PROJECT-4 REPORT

- This project is on the implementation of Latent Dirichlet Allocation from Scratch and using the topic representation of LDA as input to logistic regression based optimization method using Newton's method
- There are two tasks here, Task 1 builds the topic model representation for the NewsGroup dataset and Task 2 uses this representation for inputs to logistic regression

TASK-1

After running the model for 500 iterations, we got the following topics for artificial and Newsgroup dataset

Artificial:

Topic 0	bank	water	river
Topic 1	loan	dollars	bank

For Newsgroup:

Topic 0	hst	shuttle	mission	pat	design
Topic 1	even	time	good	large	high
Topic 2	bill	moon	support	people	blah
Topic 3	space	science	internet	world	sci
Topic 4	car	clutch	shifter	sho	mph
Topic 5	power	question	long	engines	such
Topic 6	edu	writes	article	apr	eliot
Topic 7	manual	cars	speed	shift	find
Topic 8	edu	gif	uci	ics	incoming
Topic 9	cars	book	car	dealer	price
Topic 10	car	ford	mustang	probe	rear
Topic 11	don	engine	driving	low	toyota
Topic 12	etc	day	point	mass	saturn
Topic 13	oil	engine	turbo	come	service
Topic 14	sky	earth	temperature	things	life
Topic 15	edu	writes	insurance	article	uiuc

Topic 16	henry	edu	toronto	spencer	writes
Topic 17	don	two	money	make	people
Topic 18	nasa	gov	mission	orbit	spacecraft
Topic 19	space	station	option	shuttle	nasa

As we can see from the above topics, the model was able to group the words into relevant topics. For topic 18, the words are related to space research. But there are some words like blah that are also included in the topics, and it does not have a semantic meaning. But the word classification to the respective topics is justifiable and the model has done a good job both on Newsgroup and Artificial data.

TASK 2

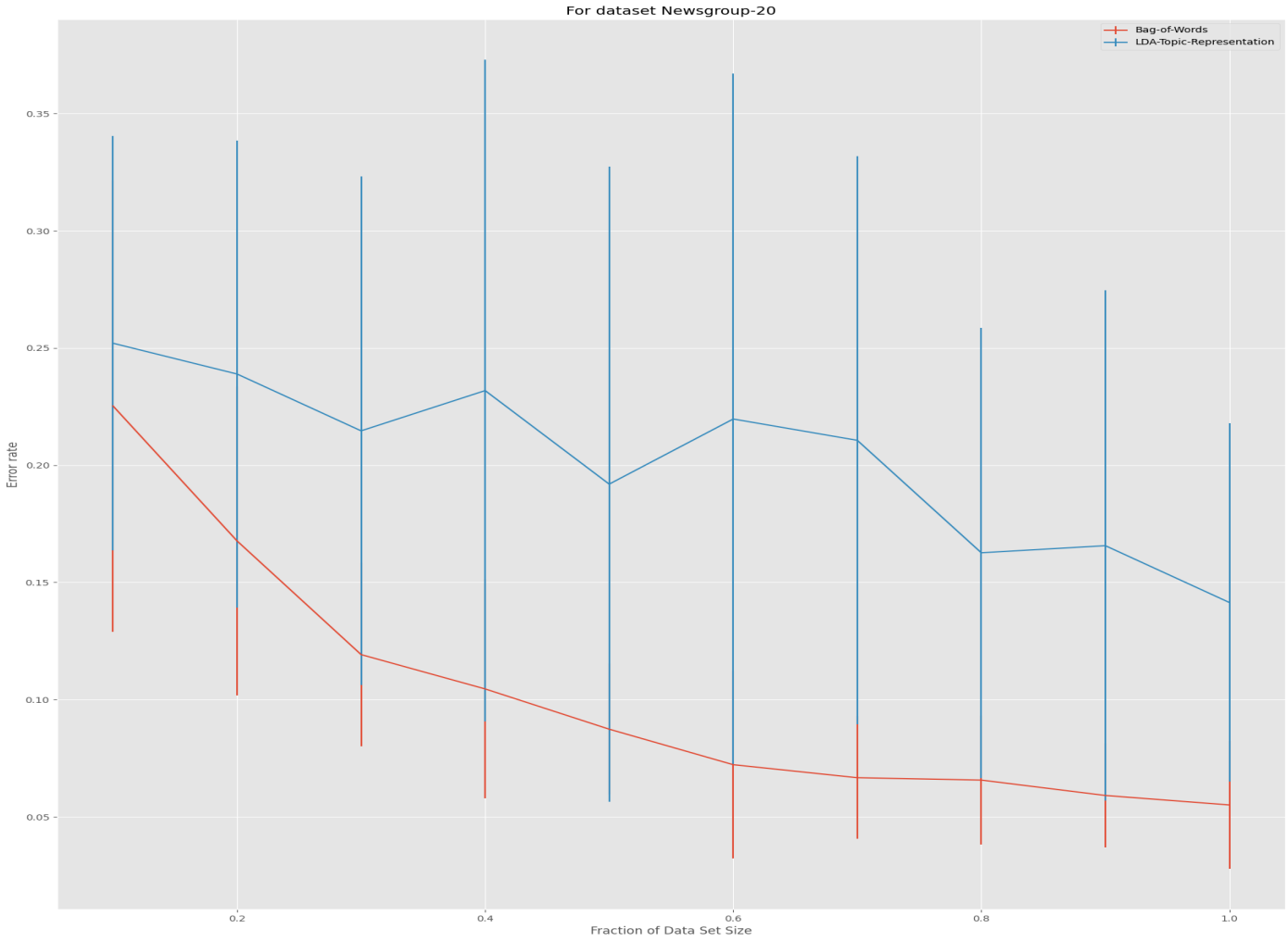
For this task, we took the document to topic representation, which is a 200x20 matrix, and the bag of words matrix, which 200x495 matrix. So the former has a smaller input size and the later has a larger input size. So the running times of the later algorithm (bag-of-Words) algorithm is significantly

higher than the running time of the LDA topic-representation algorithm (nearly 50 times)

```
BOW Model took  245.75354194641113  seconds
Topic-LDA Model took  5.784987211227417  seconds
```

But the LDA topic representation took around 4077 seconds. So it took a lot of time to get the topic represented.

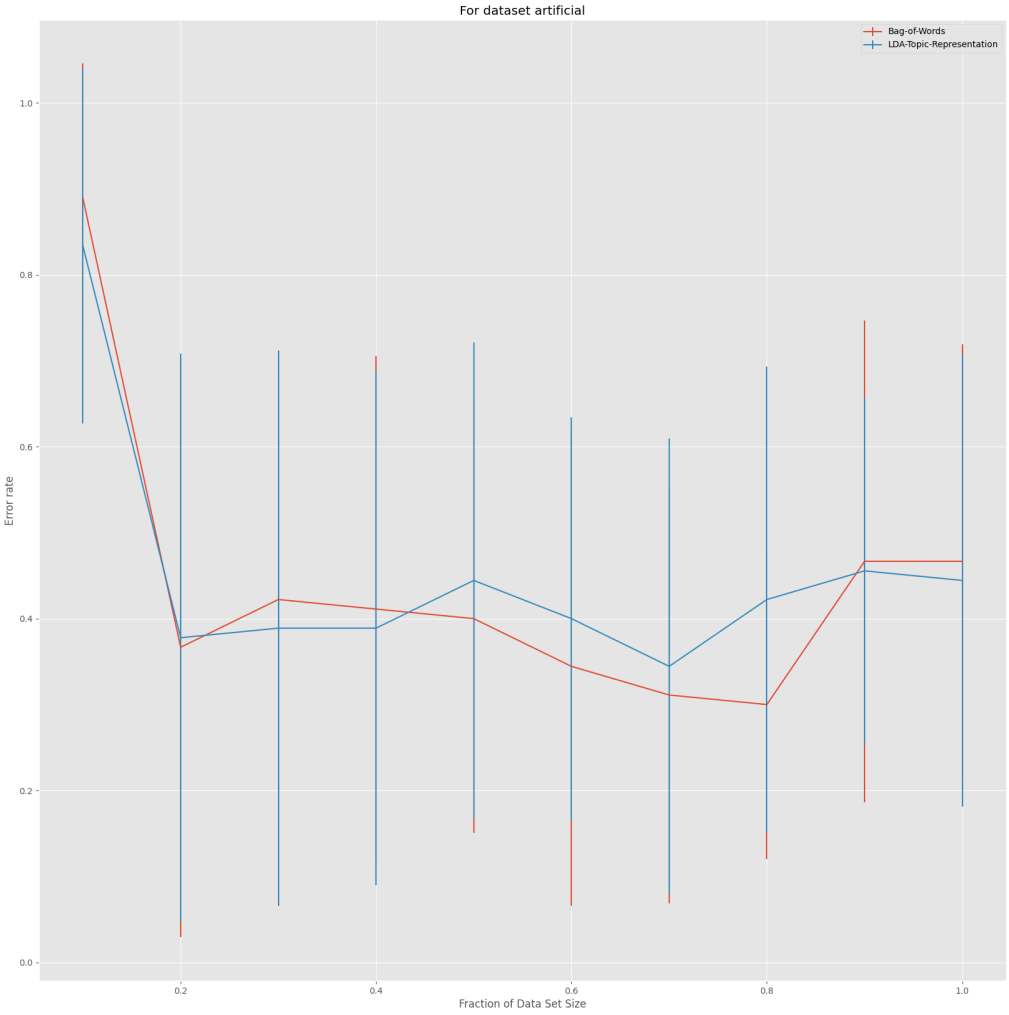
Similarly, for error rate, we have the following plot



We can see that the Bag-of-Words model has a lower error rate as compared to the Topic-LDA model. But it took a lot more time to run as compared to LDA and if we compare the ratio of the error rate to the time is taken, the LDA topic model is a bit superior. Also, the error rate between these two models is comparable, and given that LDA

achieved it with nearly 48X less time, means that LDA has accurately reduced the input feature space. Also, the reduced feature space is nearly 25 times less, so yes, LDA was able to reduce the feature dimension. However, the time required to generate the topic model is substantially high as compared to the total running time of the BoW algorithm.

For artificial


For dataset artificial
Bag-of-Words
LDA-Topic-Representation
Error rate
Fraction of Data Set Size
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.4
0.6
0.8
1.0