Task 1: Gibbs Sampling

Results:

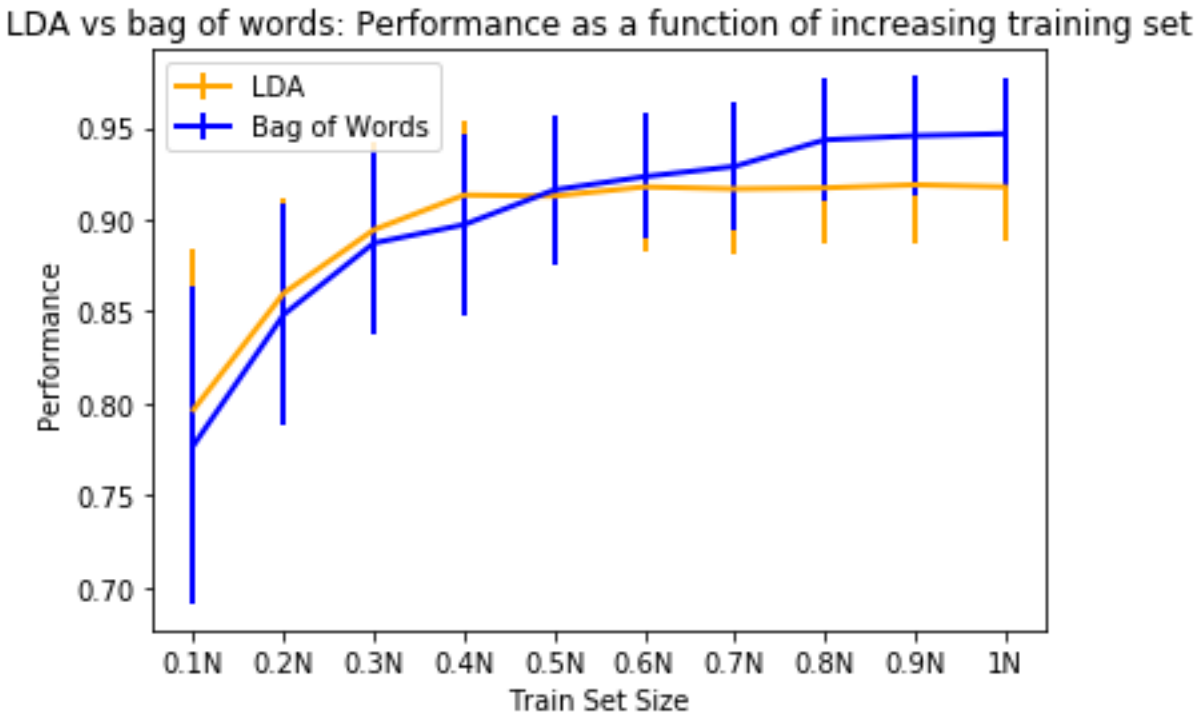
1	good, people, money, place, case
2	oil, service, lights, bmw, drive
3	science, edu, internet, information, resources
4	station, shuttle, launch, option, redesign
5	such, mars, mission, propulsion, ship
6	henry, edu, toronto, spencer, writes
7	bill, moon, earth, support, day
8	car, ford, nice, probe, george
9	hst, mission, pat, access, arrays
10	edu, writes, article, apr, don
11	time, speed, used, good, large
12	cars, don, heard, diesels, etc
13	don, find, even, make, extra
14	engine, sho, power, turbo, toyota
15	space, nasa, gov, long, sci
16	sky, light, rights, night, things
17	car, clutch, shifter, manual, shift
18	insurance, geico, make, two, true
19	edu, gif, uci, ics, incoming
20	part, spacecraft, oort, system, book

Do the topics obtained make sense for the dataset?

Yes, they do make a sense because the word grouped in a topic belongs/corresponds to specific context. For e.g., topic 2 is related to automation, topic 3 is related to information, topic 14 talks about engines, etc. There are some rows which have words not related to topic. This can be accounted to the dimensionality reduction of LDA which leads to misclassification of words to topic.

Task 2: Classification

Plots:



Observations:

1. When the data is very less then standard deviation is also more, and accuracy is less. As the data increases standard deviations goes down and accuracy increases.
2. Dimensionality reduction is done in LDA, so data is lost. Whereas in bag of words, data isn't lost. Hence, as the training size increases, bag of words performs better than LDA.
3. LDA is better for small datasets whereas bag of words become better for larger datasets.