B555 – Machine Learning

Submitted by : Nikita Bafna

Assignment 3: Generalized linear model using Logistic Regression

Introduction:

In this experiment, we implement the collapsed Gibbs sampler for LDA inference, and compare the LDA topic representation to a “bag-of-words” representation with respect to how well they support document classification.

Below are the parameters set in the experiment:

- Number of topics K = 20
- Dirichlet parameter for topic distribution α = 5/K
- Dirichlet parameter for word distribution β = 0.01
- number of iterations to run sampler Niters = 500

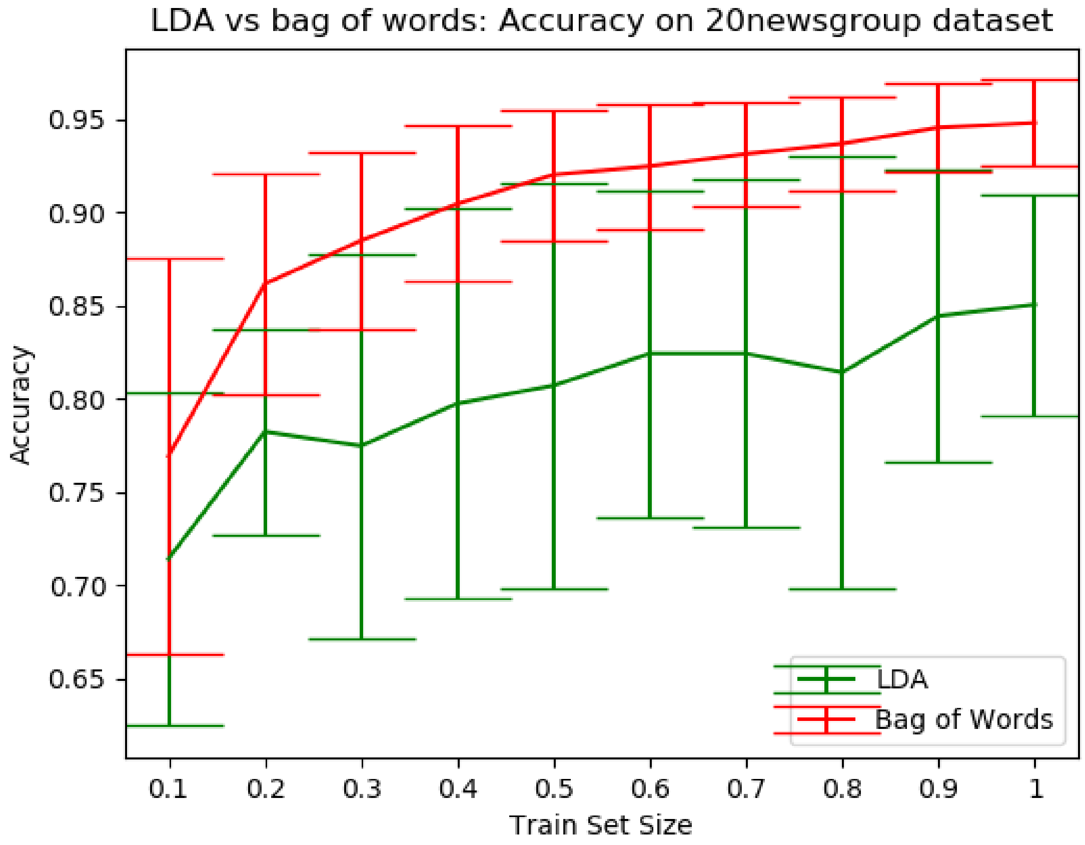
We get below 5 frequent words for the topics:

0	Space	shuttle	cost	program	nasa
1	Writes	article	edu	use	apr
2	Space	idea	history	book	world
3	Oil	service	blah	change	changing
4	Station	redesign	option	team	capability
5	Insurance	geico	want	companies	whether
6	Edu	writes	article	mustang	info
7	Mission	hst	pat	mass	solar
8	Car	shifter	sho	clutch	ford
9	Henry	spencer	toronto	george	zoo
10	Large	don	another	good	called
11	Etc	back	day	bill	sun
12	System	part	spacecraft	Each	detectors
13	Diesels	torque	emissions	Nothing	heard
14	Time	point	great	Lights	extended
15	Make	even	never	Doesn	interested
16	Engine	toyota	once	Feel	seat
17	Edu	gif	uci	Ics	incoming
18	Science	internet	information	technology	space
19	Don	cars	problem	transmission	manual

Observation:

The results obtained by LDA do make sense. This can be observed particularly in topic 0 (space, shuttle, cost, program, nasa) which is about space and topic 8 (car, shifter, sho, clutch, ford) which is about cars. Most of the topics have some related words.

Below is the graph we got for LDA vs Bag of Word model:



Observation:

It seems that when the data size is very small both BOW and LDA have almost similar accuracy. But as the size of the data increases BOW has significantly better accuracy compared to LDA. **It is also observed that the graph varies each time with the run and has different accuracy rate for LDA model for different train size.**

There could be various reasons for it:

1. Maybe LDA hasn't converged in 500 iterations.
2. Maybe K = 20 isn't the optimal value for the no of topics.
3. May the parameters set for alpha and beta aren't optimal.

4. Or maybe BOW provides a richer representation compared to LDA for a smaller dataset. For larger documents with lots of topics maybe LDA could work better.