Christian Zinck

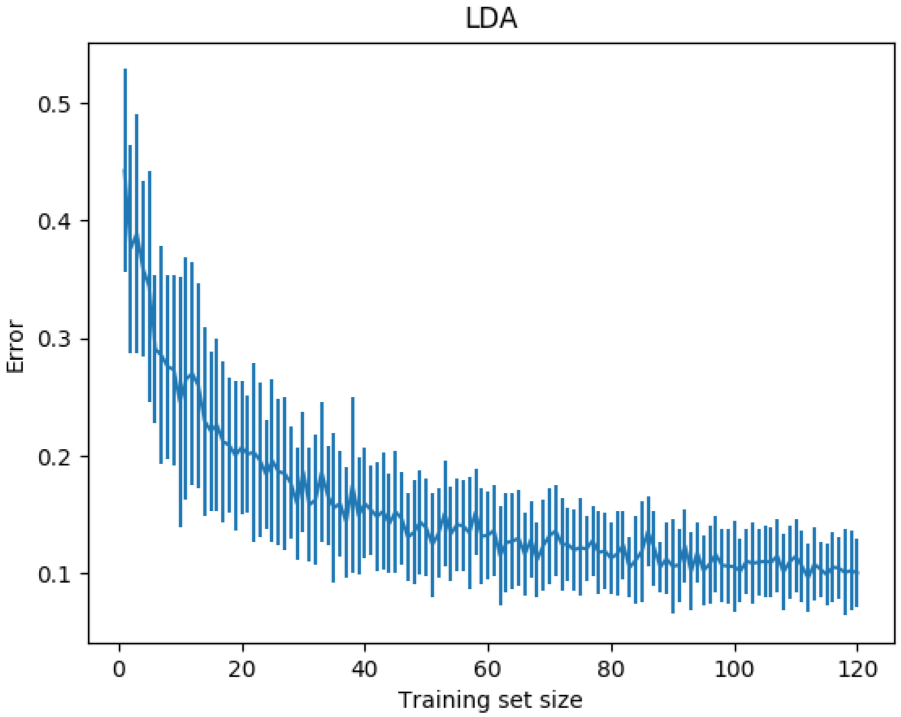
PP4

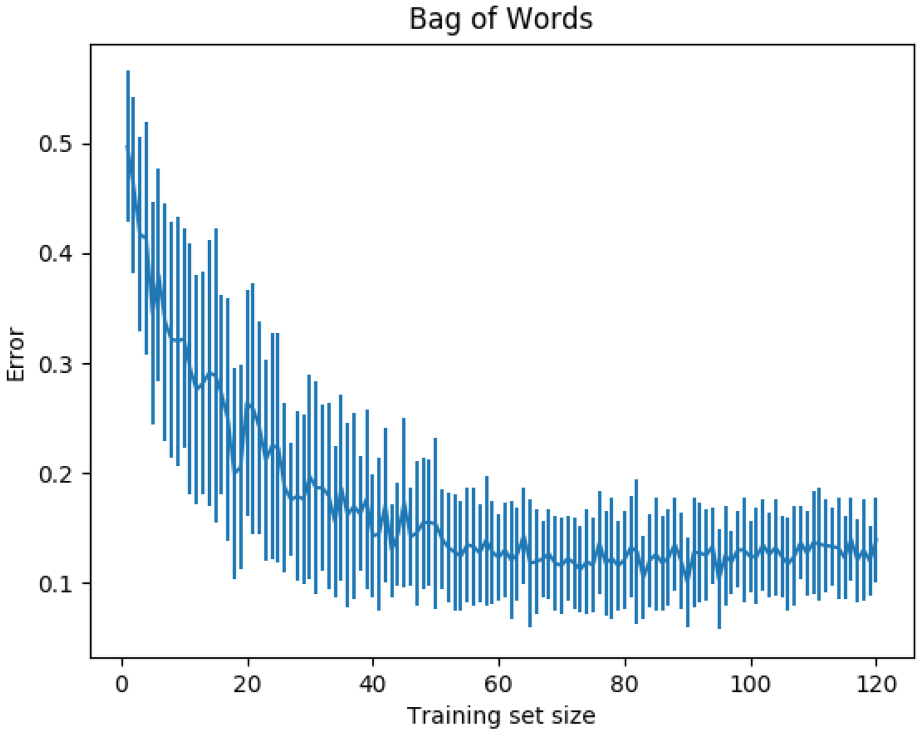
Task 1:

```
edu,article,writes,apr,good
space,bill,time,power,light
don,make,use,want,mph
george,uiuc,resources,howell,back
edu,insurance,gif,uci,ics
cost,spacecraft,nasa,another,second
ford,engine,turbo,mustang,heard
sky,edu,temperature,writes,large
hst,pat,mission,access,net
space,station,nasa,option,science
cars,two,manual,toyota,speed
henry,edu,toronto,spencer,writes
idea,book,money,want,blue
mission,solar,mars,design,arrays
car,clutch,shifter,sho,shift
car,price,power,dealer,small
edu,article,writes,internet,don
shuttle,space,diesels,missions,work
oil,engine,service,come,change
launch,day,low,mass,temp
```

The 5 most frequent words from most to least common in each of the 20 news topics are shown above. A quick inspection of the words reveals that the topics are well formed and mostly consist of automobile and spacecraft terminology, which are the 2 classes in the classification task. There is a great deal of overlap between the topics, most likely because there are not 20 unique topics represented in the corpus and the total number of words in the vocabulary is only around 400.

Task 2:


LDA


Bag of Words

The learning curves for Latent Dirichlet Allocation with Bayesian Linear Regression and the unigram Bag of Words model are very similar. The descent of the LDA learning curve is slightly smoother because each additional document will only slightly alter the topics produced by Gibb's sampling, while the addition of a document in the BoW model may change the probabilities more drastically. The descent of LDA is also slightly steeper due to the fact that a valid topic can be discerned from relatively few documents, while an approach that relies on word frequency will require a larger training set to achieve the same accuracy. The standard deviations appear to be consistently smaller for the LDA model, but I cannot come up with an explanation for why that is the case. Ultimately, as the training set size reaches the maximum (60%), both models approach approximately 10% test error on average, with LDA slightly outperforming BoW.