Report for Project 4

Zhaokun Xue

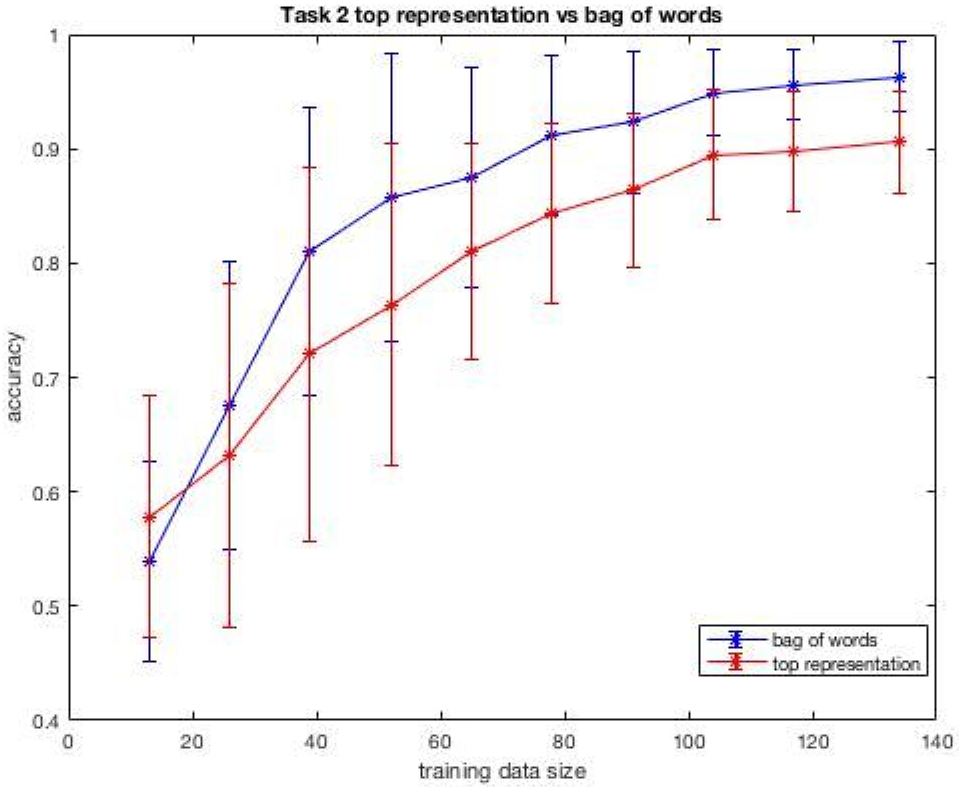
- **Task1**

Results

Topic 1	station,shuttle,launch,option,space
Topic 2	insurance,geico,mail,quote,people
Topic 3	engine,small,power,turbo,driving
Topic 4	car,ford,find,feel,probe
Topic 5	clutch,shifter,manual,sho,shift
Topic 6	article,edu,writes,apr,called
Topic 7	space,such,time,long,sci
Topic 8	cars,diesels,put,lot,heard
Topic 9	nasa,science,internet,information,mars
Topic 10	sky,people,light,money,rights
Topic 11	mission,hst,solar,shuttle,pat
Topic 12	etc,earth,large,life,planets
Topic 13	writes,good,system,oort,cloud
Topic 14	edu,writes,apr,article,eliot
Topic 15	henry,toronto,spencer,writes,zoo
Topic 16	edu,gif,uci,ics,incoming
Topic 17	oil,bill,service,moon,back
Topic 18	want,even,two,cost,extra
Topic 19	cars,george,mustang,bit,big
Topic 20	don,make,book,price,use

- **Task2**

  Plot



○ Discussion your observations on the results obtained:

Based on the results I got, in general, the “bag of words” method has higher accuracy than “topic representation” method, and “bag of words” also has smaller covariances. “topic representation” beats “bag of words” only at training size 13. According to my results, the dimension for “bag of words” is 405 which is the vocabulary size, and the dimension for “topic representation” is 20 which is the number of topics. It is much easier to find a separate hyperplane in higher dimension. That is why we have better accuracy for “bag of words”. However, it takes much less time to run the “topic representation” method, because “topic representation” method has fewer features, which makes the computation time of

Hessian for Newton method much shorter. On the other hand, the feature space of “bag of words” would be really large which could lead the computation time for Newton method really long.