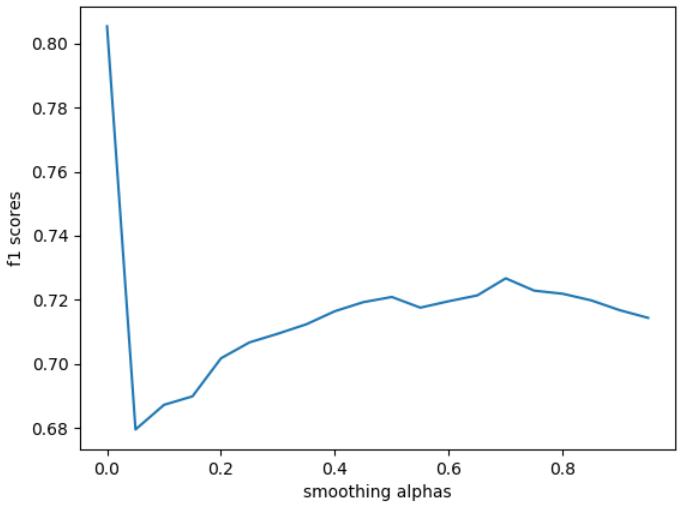
SI 630 - homework 1
Kaggle username: Yixi Li
Q1 Naive Bayes

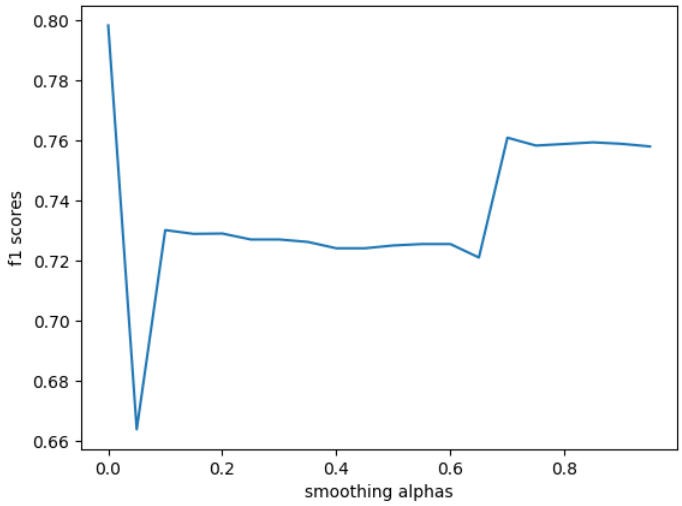
1. The best F1 score on development data is 0.8053 with smoothing alpha equals 0.

2. The model doesn't work very well when smoothing alpha equals 1.0 compared with smaller values.



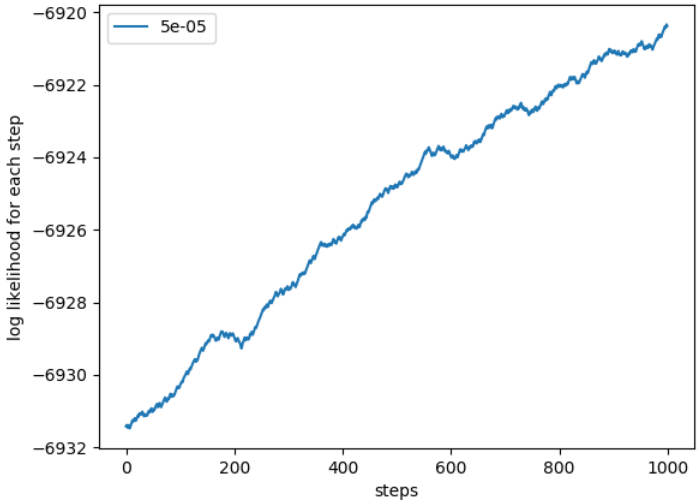
3. For better_tokenizer, I improved the following aspects:
a. split the phrases by special characters except for '\" with regex.
b. remove those special characters.
c. change all the character to lower cases.
I did these because I found some of the phrases contains special characters and both upper and lower cases existed, however, I would not want to split phrases such as 'I'm', 'you're', etc.

4. The better_tokenizer decreased the f1 score when smoothing alpha is smaller than 0.1 but increased the f1 scores when smoothing alpha is larger than 0.1.

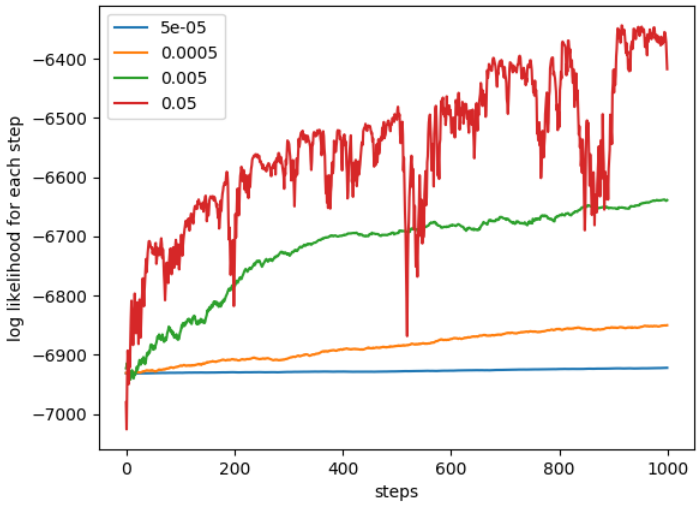


Q2 Logistic Regression

1. No, the model hasn't converged when learning rate = 5e-05 and num_step = 1000. It may take more steps to converge.



2. As the learning rate gets larger, the converging speed gets quicker.



3. The best F1 score is 0.7903 when learning rate = 5e-02 and num_step = 2e06.