Hw1 report

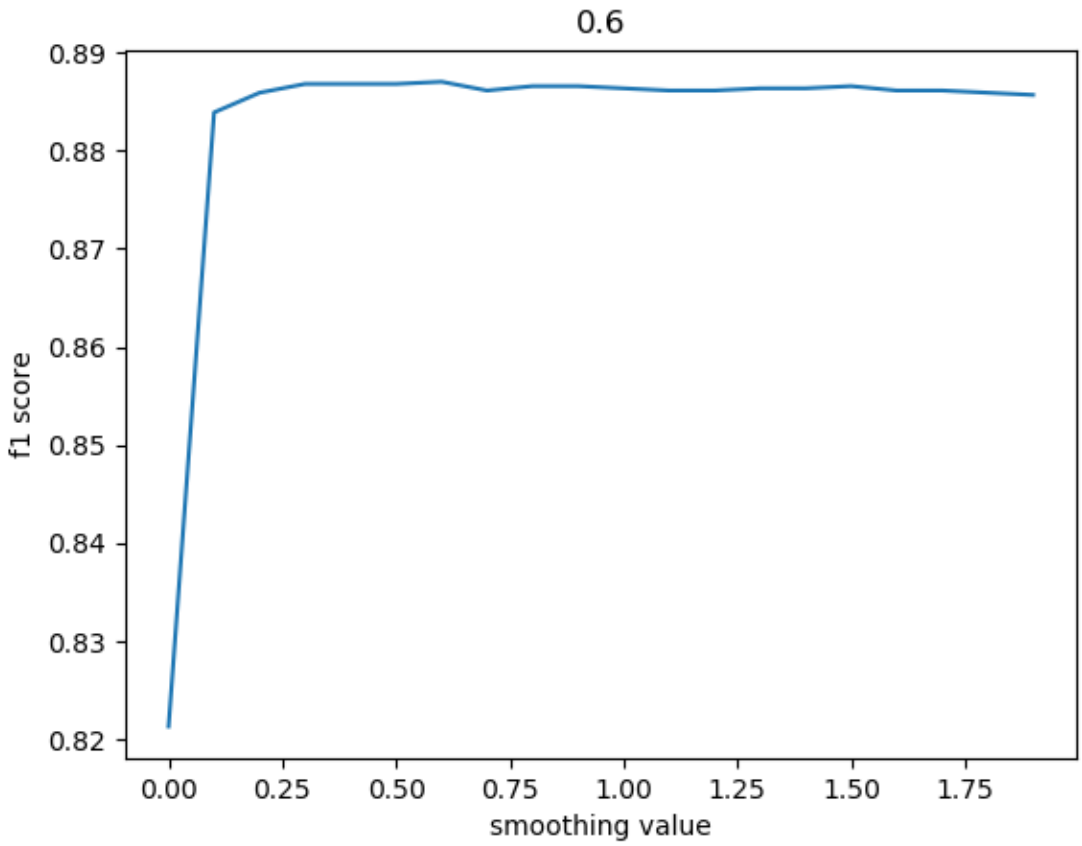
Bicheng Xu

NB

Part1:

F1 on no smoothing: `0.812639571237`

`F1 – smoothing a (when using better_tokenize()):`



When a = 0.6, the f1 is at its peak. When a = 1, it still gives a good performance.
So I will choose a = 0.6 as my best model.

Part 2

For better tokenization, I did:

- Remove @people

- Remove url
- Remove &#number
- Remove non-charactors both at the beginning and end of a word
- Use ‘encoding’ as ‘utf-8’ to include emojis

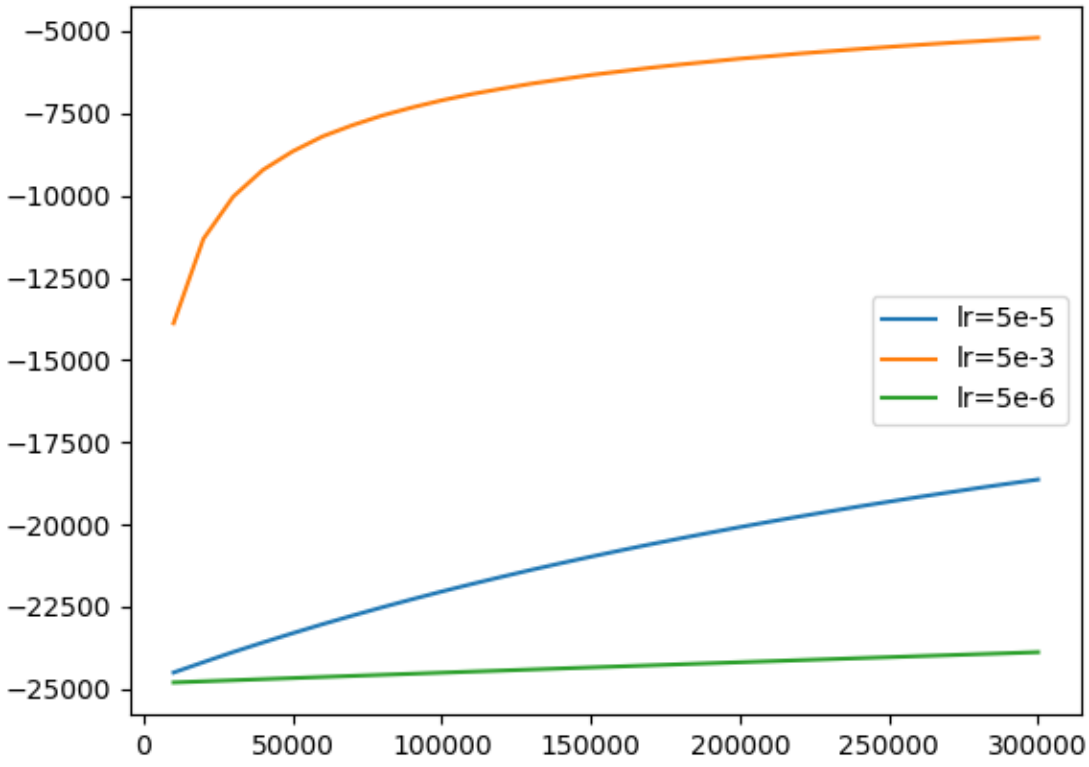
Comparison of F1 (when a = 1):

Tokenize: 0.869361322019
Better_tokenzie: 0.886333184457
Therefore, better tokenization function gives me a higher F1 score.

Logistic Regression

F1: 0.53394372487717734

`Comparing learning-rate:`



As can seen, when learning rate is 5e-3, it converges more quickly. When the learning rate is 5e-5 or 5e-6, it's rather slow. So I will choose 5e-3, even though It hasn't converged after 30,000 steps.

After I set the number to 40,000, the F1 begins to converge, but still low: 0.575703439035