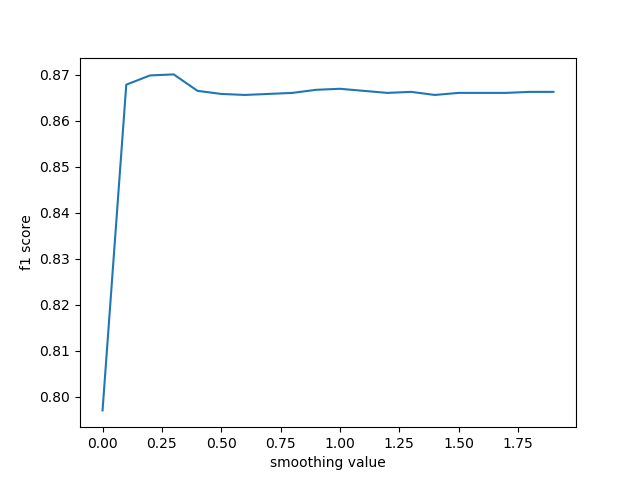
Hw1 report

NB

Part1:

F1 on no smoothing: 0.812639571237

F1 – smoothing a (when using tokenize()):



When a = 0.3, the f1 is at its peak. When a = 1, it still gives a good performance.

So I will choose a = 0.3 as my best model.

Part 2

For better tokenization, I did:

* Make all characters lower in case, so ‘happy’ and ‘Happy’ will be the same.
* Use ‘encoding’ as ‘utf-8’ to include emojis
* Strip out special characters at the beginning and end of a word. So ‘happy’ and ‘happy”’will be the same.
* Also include meaningful punctuations such as ”!”, “…” and “?”.

Comparison of F1 (when a = 1):

Tokenize: 0.869361322019

Better\_tokenzie: 0.870477891916

Therefore, better tokenization function gives me a higher F1 score.