COMP9318 Project Report

Bonus Part

Yuhan He (z5224997)
Zihao Xu (z5184152)

Implementation details of the model

In order to reduce the *Total Number of Incorrect Labels* in Q1, we choose Absolute Discounting instead of Add-1 smoothing to smooth the emission probabilities.
There is evidence[1] that Absolute Discounting performs better than Add-1 smoothing. We observed that lots of incorrect labelling are caused by the unknown token. Absolute Discounting treats unknown tokens in a more flexible way than Add-1 smoothing.
Now, we declare the meaning of symbols we used in the calculation:
M : The total number of symbols appear in the Symbol_File

F : A list with length of the number of states, $F[i]$ is the number of different symbols transmitted from the i-th state.

N : A matrix in the shape of *(total states number) * (total symbol number + 1)*. We added one more symbol: *UNK* to denote all tokens that have not to appear in the *Symbol_File*. The symbol ID of *UNK* is the total symbol number M.

$N[i][j]$: The frequency for each symbol i.

S : A list with length of the number of states, $S[i]$ is the total frequency number of all possible symbols for the i-th state.

We choose an arbitrary value $a[i] = \frac{1}{(F[i]+S[i])}$

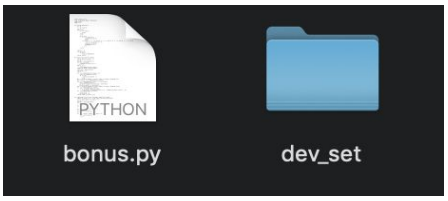
The formula we used to calculate the probability that from state i to symbol j as follows:

$$if\ N[i][j] = 0,\ B[i, j] = \frac{F[i]*a[i]}{M - F[i] + 1}$$
$$else,\ B[i, j] = \frac{N[i][j]}{S[i]} - a[i]$$

Absolute discounting is a serendipitously discovered estimator, it corrects empirical frequencies by subtracting a constant from observed categories, which it then redistributes among the unobserved. It outperforms classical estimators empirically and has been used extensively in natural language modelling.

Instruction of how to execute the code:

1. Import *bonus*
2. Call the defined function:
   `absolute_discounting_decoding(State_File, Symbol_File, Query_File)`
   The parameters are the same which we used in *Q1.*




```
$ python
Python 3.6.3 |Anaconda, Inc.| (default, Dec  5 2017, 17:30:25)
[GCC 4.2.1 Compatible Clang 4.0.1 (tags/RELEASE_401/final)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
>>> import bonus as b
>>> r = b.absolute_discounting_decoding('./dev_set/State_File','./dev_set/Symbol_File', './dev_set/Query_File')
>>> for ele in r:
...     print(ele)
...
[24, 0, 1, 2, 3, 18, 4, 18, 5, 6, 25, -59.36829471638931]
[24, 2, 3, 18, 4, 18, 5, 6, 25, -37.89136202707897]
[24, 8, 9, 18, 4, 18, 5, 6, 25, -56.91496333289705]
[24, 8, 9, 7, 1, 2, 3, 18, 4, 18, 5, 6, 25, -76.88916925002172]
[24, 8, 9, 2, 2, 2, 0, 23, 0, 19, 1, 2, 18, 4, 18, 5, 6, 25, -129.20860914231554]
[24, 0, 19, 2, 3, 18, 4, 18, 5, 6, 25, -52.112970649544046]
[24, 0, 19, 1, 2, 2, 3, 18, 4, 18, 5, 6, 25, -61.26789350732184]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -49.235000878660166]
[24, 0, 19, 1, 2, 3, 18, 4, 4, 18, 5, 6, 25, -61.00651985489211]
[24, 8, 8, 9, 18, 2, 3, 4, 18, 4, 18, 5, 6, 25, -81.52255794046877]
[24, 1, 2, 3, 18, 4, 18, 5, 6, 25, -42.45301096826461]
[24, 0, 19, 1, 2, 3, 18, 4, 4, 18, 5, 6, 25, -60.279383198153816]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -52.87078368520328]
[24, 2, 3, 18, 4, 18, 5, 6, 25, -49.49431176482504]
[24, 0, 18, 1, 20, 1, 2, 3, 18, 4, 18, 5, 6, 25, -61.68514602149918]
[24, 8, 9, 15, 9, 18, 1, 20, 1, 2, 3, 18, 4, 18, 5, 6, 25, -79.06215171325985]
[24, 0, 18, 4, 9, 18, 2, 3, 18, 4, 18, 5, 6, 25, -78.28303835107228]
[24, 0, 19, 1, 2, 2, 3, 18, 4, 4, 18, 5, 6, 25, -80.88209450829075]
[24, 1, 2, 2, 3, 18, 4, 18, 5, 6, 25, -62.35251300720911]
[24, 7, 7, 2, 3, 18, 4, 18, 5, 6, 25, -56.402993162540724]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -49.94294473808129]
[24, 8, 9, 18, 4, 18, 5, 6, 25, -55.085261277626124]
[24, 0, 19, 1, 2, 3, 18, 4, 4, 18, 5, 6, 25, -60.767285304601 78]
[24, 0, 19, 1, 2, 3, 18, 4, 4, 18, 5, 6, 25, -53.88078272117967]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -47.94676676529601]
[24, 0, 19, 1, 2, 3, 18, 4, 4, 18, 5, 6, 25, -57.63554697214787]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -50.315695607605406]
[24, 8, 9, 1, 18, 4, 18, 5, 6, 25, -57.902743112342115]
[24, 0, 1, 2, 2, 2, 3, 18, 4, 18, 5, 6, 25, -82.51475615513445]
[24, 2, 3, 18, 4, 18, 5, 6, 25, -40.71643801419017]
[24, 2, 3, 18, 4, 18, 5, 6, 25, -38.27615292156772]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -49.82824297427295]
[24, 0, 19, 1, 2, 2, 3, 18, 4, 18, 5, 6, 25, -60.43749558861376]
[24, 8, 8, 9, 18, 4, 18, 5, 6, 25, -53.284329254945604]
[24, 16, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -54.47260507748032]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -53.82257111098554]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -53.6834023642769]
[24, 0, 19, 1, 2, 3, 18, 4, 18, 5, 6, 25, -52.57166825279979]
```

Reference:

[1]Boodidhi, S., 2011. Using smoothing techniques to improve the performance of Hidden Markov's Model.https://digitalscholarship.unlv.edu/cgi/viewcontent.cgi?referer=https://scholar.google.com.au/&httpsredir=1&article=2008&context=thesesdissertations