NPFL097 Assignment 2
Chinese Segmentation

Andrew McIsaac

December 10, 2021

1. Implement the model based on Chinese Restaurant Process as described in the previous lecture. Set the hyperparameters $\alpha = 100$, $p_c = 0.5$, $p_{cont} = 0.99$, $T = 1$. *(4 pts)*

   See `crp.py`

2. Debug your code, check the output segmentation and try to change the parameters to obtain better segmentations. *(2 pts)*

   See `crp.py`

3. Download the gold data and the evaluation script. What precision and recall do you get? `(1 pt)`

   Default hyperparameters: $\alpha = 100$, $p_c = 0.5$, $p_{cont} = 0.99$, $T = 1$, iterations = 100

   Results for default hyperparameters are Precision=0.172, Recall=0.287, F1 = 0.215

Table 1: Different numbers of iterations with otherwise default hyperparameters

	20	50	100	200
Precision	0.141	0.169	0.172	0.181
Recall	0.237	0.281	0.287	0.302
F1	0.177	0.211	0.215	0.227

Table 2: Different values of α with otherwise default hyperparameters

	1	20	100	200
Precision	0.176	0.171	0.172	0.180
Recall	0.296	0.287	0.287	0.299
F1	0.221	0.215	0.215	0.225

Table 3: Different values of p_c with otherwise default hyperparameters

	0.3	0.5	0.7
Precision	0.176	0.172	0.176
Recall	0.293	0.287	0.291
F1	0.220	0.215	0.219

Table 4: Different values of p_{cont} with otherwise default hyperparameters

	0.8	0.99	0.999
Precision	0.167	0.172	0.168
Recall	0.272	0.287	0.282
F1	0.207	0.215	0.211

4. Try to do annealing and run the model for different temperatures. You can also try changing the temperature during the sampling. E.g., start with a higher temperature and then gradually decrease to zero. *(2 pts)*

Table 5: Different temperatures, except for first column, which decays from a temperature of 10 by multiplying T by 0.95 every iteration until it reaches 1

	$10 \rightarrow 1$	0.5	1	2
Precision	0.202	0.097	0.172	0.135
Recall	0.327	0.159	0.287	0.269
F1	0.250	0.121	0.215	0.180

Having a form of decay by starting with a high T and gradually decreasing results in the best performance increase.

5. Instead of Chinese Restaurant Process, try to employ the Pitman-Yor Process. Does it improve your results? *(1 pt)*

   Yes, with Pitman-Yor process and otherwise default parameters, $d = 0.5$ improves the performance the most (F1 Chinese Restaurant Process was 0.215 vs 0.232 for Pitman-Yor Process).

Table 6: Pitman-Yor Process for different values of d

	0.2	0.5	0.9
Precision	0.175	0.186	0.182
Recall	0.290	0.307	0.309
F1	0.219	0.232	0.229

Combining best results ($\alpha = 200, p_c = 0.3, p_{cont} = 0.99, \text{iterations} = 100, annealing = True, d = 0.5$), and with T going from 3 to a minimum of 0.3 gives Precision = 0.236, Recall = 0.361, F1 = 0.285.

The segmented data is provided in `data_small_segmented.txt`.