NPFL097 Assignment 2
Unsupervised Segmentation of Text

David Mareček

November 2024

In this assignment, you will get a document with English news with all the spaces removed. Your task is to implement the unsupervised text segmentation model using Chinese Restaurant Process as presented in the previous lectures and tune its hyperparameters to obtain as good result as possible. The preferred programming language is Python.

Send me your final source code, the output with the best obtained segmentation, and the best setting of parameters by email: `marecek@ufal.mff.cuni.cz`.

Tasks:

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)*

2. Debug your code, check the output segmentation and try to change the parameters to obtain better segmentations. Test also very high values of parameter α (e.g. 10000) and parameter p_c (e.g. 0.99). *(2pts)*

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

4. Implement the annealing and run the model for different temperatures. Experiment also with changing the temperature during the sampling. E.g., start with a higher temperature so that the algorithm searches the whole space well, and then, gradually decrease it, so that the algorithm converges. *(2pts)*

5. Instead of the Chinese Restaurant Process, employ the Pitman-Yor process supporting longer tails. Test several discount parameters. Does it improve your results? *(1 pt)*