Small Problem 5: PCFG Sentence Completion

Introduction

In this problem, you are given a probabilistic context free grammar G and a prefix x of a string generated by that grammar.

Your goal is to evaluate the conditional probability that the grammar will generate the complete string xy for a given suffix y: $P(xy|x, G)$.

Grammar: Note that this grammar produces strings of unbounded length, but produces a finite string with probability 1.

S → AB (0.25)	A → a (0.05)	B → b (0.5)	C → d (0.35)
S → BC (0.2)	A → b (0.3)	B → c (0.3)	C → e (0.1)
S → AC (0.4)	A → S (0.65)	B → d (0.2)	C → S (0.55)
S → CA (0.15)			

Queries and Metrics

Query 1: What is the conditional probability that a string beginning with "bd" terminates with the complete string "bdcb": $P(y = \text{"bdcb"}|\text{prefix}(y) = \text{"bd"})$ where y is the complete string that is generated by S.

Query 2: What is $P(y = \text{"cddb"}|\text{prefix}(y) = \text{"cd"})$?

Metric: Square of the difference in negative log probability ("surprise") between the true and the computed conditional probability. $\left(-\log P(xy|x) + \log \hat{P}(xy|x)\right)^2$.

Submission:

The metric value should be computed for each elapsed time step (by calling the provided code or by implementing yourself). The metric value should be reported for several elapsed time steps. The number of elapsed time steps should be sufficient to establish an "informative profile".

For further details regarding submission of the metric and your code, please refer to the main CP4 problem description document, e.g. PPAML-Challenge-Problem-4.pdf.

Sample output for this problem has been provided in the "sampleoutput" folder:

`problem-5-query-1-metric-1.csv`

`problem-5-query-2-metric-1.csv`

Further details on this problem can be found in the provided sample solution:
`ppaml-cp4/solutions/problem5`