AI611μ Word Prediction with N-Grams Model using Python

Quizz 2: Bigram model training

This assessment evaluates the following competencies:

- *AI201 – Train an N-Grams model from a given text corpus* (+1)
- *AI102 – Formally describe N-Grams models thanks to probabilities* (+1)
- *AI501 – Write an application that solves the word prediction problem with N-Grams models* (+1)

For the two first assessed competencies (AI201 and AI102), you have to complete the bigram model of the following corpus with six sentences: *"I like to eat Chinese food. Chinese people like to eat noodles. You like to drink a lot. I drink a lot of water. Chinese people drink water. I eat Chinese noodles."*. Find the two missing values and detail your calculation:

	i	like	to	eat	chinese	food	people	noodles	you	drink	a	lot	of	water
i	0	0,33	0	0,33	0	0	0	0	0	0,33	0	0	0	0
like	0	0	1	0	0	0	0	0	0	0	0	0	0	0
to	0	0	0	0,67	0	0	0	0	0	0,33	0	0	0	0
eat	0	0	0	0	0,67	0	0		0	0	0	0	0	0
chinese	0	0	0	0	0	0,25	0,5	0,25	0	0	0	0	0	0
food	0	0	0	0	0	0	0	0	0	0	0	0	0	0
people	0		0	0	0	0	0	0	0	0,5	0	0	0	0
noodles	0	0	0	0	0	0	0	0	0	0	0	0	0	0
you	0	1	0	0	0	0	0	0	0	0	0	0	0	0
drink	0	0	0	0	0	0	0	0	0	0	0,67	0	0	0,33
a	0	0	0	0	0	0	0	0	0	0	0	1	0	0
lot	0	0	0	0	0	0	0	0	0	0	0	0	0,5	0
of	0	0	0	0	0	0	0	0	0	0	0	0	0	1
water	0	0	0	0	0	0	0	0	0	0	0	0	0	0

For the last assessed competency (AI501), compute the following probability and detail your calculation, according to the trained model: $P(\texttt{<s>you like to eat chinese people</s>}) =$