Tutorial 3: N-gram and Language Model

Q1. Given the following three word sequences (i.e., the corpus).

very good tennis player in US Open

tennis player US Open

tennis player qualify play US Open

(i) Build a table of bigram counts from the word sequences.
(ii) Compute the bigram probabilities using Laplace smoothing.

Q2. Write out the equation for trigram probability estimation, and use the equation to compute the trigram probability for $P(US|\ tennis\ player)$ and $P(player\ |\ good\ tennis)$ according to the corpus given in Q1.

Q3. Given the bigram probability in the following table, compute the probability of "I eat Chinese food". Explain how you compute the probability. State your assumptions and if more probability values are needed, you may use random values.

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Q4. Why do we need to do smoothing for language model?

Q5. Given some text, what are the general steps to collect all counts needed for building an n-gram language model?

Q6. **For discussion only**: You are given a text collection of 100GB, and asked to train a bigram language model. You have a computer with 16GB ram and 1TB storage. Think about the best choices (steps) for implementation.