Chapter 22 | Questions

1. How many entries does a trigram character model in a language with 100 characters have?

   Answer: 100 times 100 times 100 = 10^6

2. We can define the probability of a sequence of characters $P(c_{1:N})$ under the trigram model by first factoring with chain rule, and then using the Markov assumption. What will be the probability of the the sequence "abc" if P("a") = 0.1, P("b") = 0.2, P("c") = 0.3, P("b"| "a") = 0.4, and P("c" | "ab") = 0.5?

$$P(c_{1:N}) = \prod_{i=1}^{N} P(c_i \mid c_{1:i-1}) = \prod_{i=1}^{N} P(c_i \mid c_{i-2:i-1})$$

   Answer: P("abc") = P("a") * P("b"| "a") * P("c" | "ab")

3. Given a text we would like to determine what natural language it is written in. For each L, we build a model by counting the trigrams in a corpus of that language. This gives us P(Text | Language). How can P(Text | Language) be used to predict P(Language | Text)?

   Answer: P(L | T) = maximum of all { P(L) * P(T | L) }

4. Given a text we would like to determine what natural language it is written in. For each L, we build a model by counting the trigrams in a corpus of that language. This gives us P(Text | Language). When building an n-gram model, if we set zero probability to uncommon n-grams such as "xgz" or "ojz" what kind of error occurs? How can linear interpolation smoothing resolve this error? Give an example by showing how the probability of "zgx" may be calculated.

   Answer:
   - Generalization error occurs, i.e. if the model encounters any character sequence that has such n-grams it assigns 0 probability
   - Linear interpolation ensures that no n-grams have 0 probabilities
   - P("zgx") = w1 * P("zgx") + w2 * P("gx") + w3 * P("x") - w1, w2, and w3 are weights that can be learned

5. Suppose there are 3 characters in a language L, and we have built a unigram model. The probabilities for the 3 characters are given by the models are P("A") = 0.25, P("B") = 0.50, and P("C") = 0.25. What will be the expression to calculate the perplexity of for the sequences "AAA" and "ABA"?

   Answer:
   - Perplexity("AAA") = P("AAA")$^{-⅓}$ = (0.25 * 0.25 * 0.25)$^{-⅓}$
   - Perplexity("ABA") = P("ABA")$^{-⅓}$ = (0.25 * 0.50 * 0.25)$^{-⅓}$

6. What are the disadvantages of building larger n-gram models such as 4-gram or 5-gram word models?

   Answer:
   - The number of character sequences will be too large
   - The model will start to memorize (instead of generalizing)

7. A search query "president lincoln" is being scored against a document D. The terms "president" and "lincoln" both appear just once in D. The length of this document D is 90% of the average length of all documents in the corpus. There are 40,000 documents that contain the term "president" and there are 300 documents that contain the term "lincoln". Assume that IDF for all queries is 1, and k = 1.2 and b = 0.75. What is the BM25 score for the query against the document D?

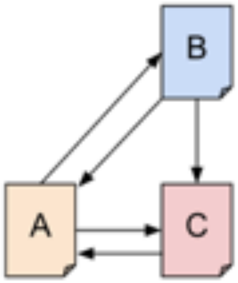
8. Given the Page rank algorithm, what do C(in_i) and N stand for?

$$PR(p) = \frac{1-d}{N} + d\sum_i \frac{PR(in_i)}{C(in_i)}$$

Answer:

- N is the total number of pages in the corpus
- C(in_i) is the count of the total number of out-links on page in_i

9. Write out the linear equations for calculating the page ranks of the pages A, B, and C in the following link network. Assume d = 0.7. You don't need to solve the system of linear equations - just list out the linear equations.



$$PR(p) = \frac{1-d}{N} + d\sum_i \frac{PR(in_i)}{C(in_i)}$$

Answer:

PR(A) = (1 – 0.7) × (1 / 3) + 0.7 × { PR(C) / 1 + PR(B) / 2 }
PR(B) = (1 – 0.7) × (1 / 3) + 0.7 × (PR(A) / 1)
PR(C) = (1 – 0.7) × (1 / 3) + 0.7 × { PR(B) / 2 + PR(A) / 2 }