1. [8 points] What is the term for …
   a. Mapping things like am/are/is to a common form like be.
   b. Separating an input stream of characters into individual words.
   c. A collection of language data (documents, texts, dialogs, etc.).
   d. A case where the model predicts something to be a member of some class, but in fact that element is not a member of that class.
   e. A simple way to specify a language in terms of symbol sequences, repetition, and alternatives.
   f. A way to map one sequence to another.
   g. Adjusting count-based estimates to reflect the fact that never is the probability of anything actually zero.
   h. A preposition plus the following noun phrase.

2. [2 pts] What is the difference between a deterministic and a non-deterministic model?
   Explain or illustrate.

3. [1] If \( P(x|a) = 0.45 \) and \( P(x|b) = 0.55 \), what is \( \text{argmax}_{c \in \{a,b\}} P(x|c) \) ?

4. [3] Sometimes the word “elephant” shows up in emails. Assume you know \( P(\text{elephant}|\text{spam}) = .0001 \) and \( P(\text{elephant}|\text{notSpam}) = .0003 \), and that a certain email contains that the word elephant. What additional information, not in this email, would you need to have in order to compute \( P(\text{notSpam}|\text{elephant}) \), that is, the probability of this email being not spam, using Bayes Rule.
5. [4] Explain the concepts “token,” “lemma,” “stem” and “morpheme” using one or more of the examples below. It may be helpful to think in terms of what preprocessing is needed to prepare documents for a sentiment classification algorithm.
   a. … in New York City …
   b. … ice cream …
   c. …uncomfortably …
   d. … written …
   e. … disGGustinG …
   f. … It’s at http://www.cs.utep.edu/nigel/slp/ ← recommended link …

6. [10] For the phrase “Denver’s mayor’s cancelled flight,” a) draw a possible parse tree using the grammar with the following rules
   - NP -> (Det) (Card) (Ord) (Quant) (AP) Nominal
   - Det -> NP ’s
   - Nominal -> Noun
   - Noun -> Denver, mayor, flight
   - AP -> Adjective
   - Adjective -> cancelled

   b) is this the only possible parse tree? That is, is this sentence syntactically unambiguous?
7. [5] True or False

   a. `re.search(r'abc', 'abd')`
   b. `re.search(r'abc', 'ABC')`
   c. `re.search(r'[ghi][mno][jlk][def]', 'gold')`
   d. `re.search(r'baa*!', 'b!')`
   e. `re.search(r'baa*!', 'baaaa!')`

8. [6] For building a spam filter, there are many features you might consider. Unigrams are an obvious choice; trigrams are another. Note at least one advantage and one disadvantage of each. Illustrate with examples from the following spam mail:

   Dear Scholar,
   
   To enhance the communication between authors and keep pace with the cutting-edge research, we sincerely invite specialists and scholars to attend the conferences which to be held in Nanjing, China in October, 2018 and in Xi’an, China in December, 2018. Once your papers are accepted, they will be published online in the cooperating journals. … The fields of the conferences range broadly like: Electrical & Electronic Engineering, Education, Mechanical Engineering, Mathematics, Computer Science, etc.
9. Most puns involve words similar in sound but very different in meaning. A common pattern is to evoke a well-known phrase using one that sounds rather similar.

For example, one can say “a bun is the lowest form of wheat” to evoke the well-known phrase “a pun is the lowest form of wit”. Sketch out how you might build a system to automatically detect puns, describing the components, how they interact, and how you would build or train them. The following examples (from Kao and Tan 2012) may help:

- Being able to fit size fourteen shoes is quite a feet. (feet → feat)
- Thieves have muscles of steal. (steal → steel)
- The magician was so mad he pulled his hare out. (hare → hair)