Things you need to know:

I. Text classification with N-grams
   • What is an “N-gram”?
   • Explain how to train an N-gram model
   • Explain how to estimate probability of a sentence using an N-gram model
   • Explain how to do “add-1 smoothing”
   • Explain in words: How to train an N-gram model to generate text (e.g., Shakespeare example from slides)

II. Text classification with Naïve Bayes
   • What is Bayes rule?
   • Explain all the terms in the Naïve Bayes classification rule:
     $$c_{MAP} = \arg \max_{c \in C} \left[ P(d \mid c)P(c) \right]$$
   • Explain how to derive the Naïve Bayes classification rule
   • Be able to use the Naïve Bayes classification rule as in the examples we did in class (you don’t need to memorize the formula above, and you won’t need a calculator during the quiz)
III. Text classification with vector space models

- Explain how to represent a document as a vector of word frequencies
- Be able to use the normalized dot product to find the distance between two documents in vector space (you won’t need to memorize the formula for normalized dot product, and again, you won’t need a calculator during the quiz)

IV. Precision/Recall

- Understand the formulas for Precision and Recall (you don’t have to memorize them), and be able to calculate these values from a confusion matrix, as in the class slides
- Know some examples where you would care more about high recall than high precision, and vice versa

V. Machine translation

- Be able to define all the terms in for statistical machine translation explained in class:

\[
Translation(T) = \arg\max_s [P(S)P(T|S)]
\]

- Be able to explain how to compute the Language Model \(P(S)\), as described in class and in Section 2 of the Brown et al. paper.

- Be able to explain (in words) the notions of “fertility” and “distortion” as described in class and in section 3 of the Brown et al. paper