CSE 4392 Special Topic: Natural Language Processing

Homework 6 - Spring 2025

Due Date: Mar 3, 2025, 11:59 p.m. Central Time

Problem 1 - 50%

For this task, you are provided with a very simple dataset consisting of labeled sequences of words, where each word is associated with a specific part-of-speech (POS) tag.

- Training Data:
  - "cats (NN) jump (VB) high (JJ)" (Noun-Verb-Adjective)
  - "apples (NN) grow (VB) red (JJ)" (Noun-Verb-Adjective)
  - "cats (NN) run (VB) quickly (RB)" (Noun-Verb-Adverb)
  - "red (JJ) dogs (NN) quickly (RB) grow (VB)" (Adjective-Noun-Adverb-Verb)
- Test Data:
  - "cats (NN) quickly (RB) jump (VB)" (Noun-Adverb-Verb)
  - "dogs (NN) grow (VB) high (JJ)" (Noun-Verb-Adjective)

Compute Training Parameters - 30%

Train the HMM by computing the training parameters (transition matrix, emission matrix and prior probabilities vector). Show steps by at least writing for each value the original fraction representing counts a/b.

Viterbi Decoding - 20%

For each test example, use Viterbi decoding to find the most probable sequence of tags. Compute the word-level and sequence-level accuracies.

Problem 2 - 40%

Implement the Hidden Markov Model in object-oriented Python. You class must have a fit method, infer method and evaluate method that you should test on the dataset shown above and match your handwritten results. Your implementation of the infer method should include the Viterbi algorithm and you should implement a bruteforce inference method for testing purposes to ensure the Viterbi algorithm is working correctly.

Preferably, submit a PDF notebook using this tool. Or submit a Python file and screenshot for the output. If your code output does not match the reported output, zero will be granted.

Problem 2 - 10%

Compare HMM and CRFs and highlight scenarios where you may use on over the other.