PoS tagging, HMM

1. Find the total number of possible hidden sequences with ‘P’ observations and ‘Q’ hidden states. (1 mark)

   Q^P

2. Consider the Viterbi sequence inference algorithm for a sequence length N with K possible states. (For POS tagging, it would be: there are N tokens and K parts-of-speech)

   Give the following answers in terms of N and K.

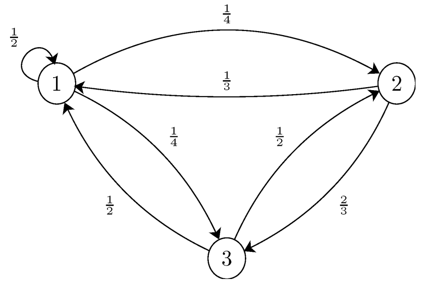
   a. What’s the space complexity of Viterbi? (1 mark)

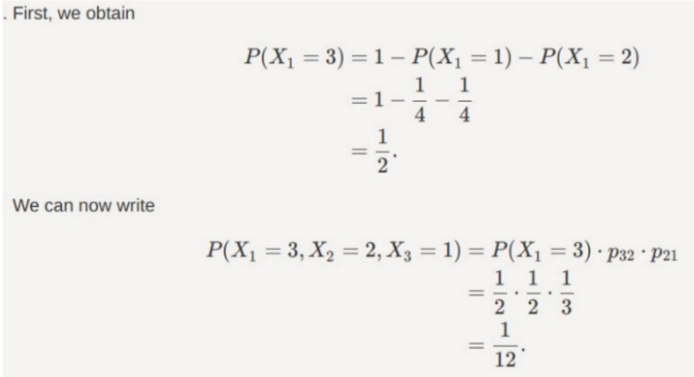
   b. What’s the time complexity of Viterbi? (1 mark)

   Time: $O(K^2N)$, because we need to do K work for each cell and K * KN = K^2N

   Space: O(KN), because we are storing a K * N sized matrix

3. Considering the markov chain in the following figure, find $P(X_1=3, X_2=2, X_3=1)$. Given $P(X_1=1) = P(X_1=2) = 1/4$ (2 marks)





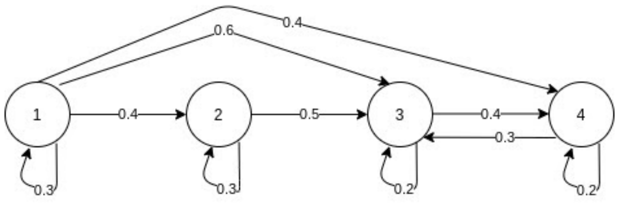
. First, we obtain

$$P(X_1 = 3) = 1 - P(X_1 = 1) - P(X_1 = 2)$$
$$= 1 - \frac{1}{4} - \frac{1}{4}$$
$$= \frac{1}{2}.$$

We can now write

$$P(X_1 = 3, X_2 = 2, X_3 = 1) = P(X_1 = 3) \cdot p_{32} \cdot p_{21}$$
$$= \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{3}$$
$$= \frac{1}{12}.$$

4. Considering the below figure, find P(X0 = 3, X1 = 4). Given P(X0 = 3) = 0.2 (1 mark)



P(X0 = 3, X1 = 4) = P(X0 = 3) * P(X1 = 4 | X0 = 3) = 0.2 * 0.4 = 0.08

5. Given a huge corpus, provide an advantage of using higher order (say, 5th order) markov chain over lower order (say, 2nd order) markov chain. (1 mark)

   Given the corpus size is large enough, higher-order (e.g., 5) markov chains will capture more context than lower-order (e.g., 2) markov chains that might help in disambiguation.

6. (a) Given a tag set T = [NN (noun), VB (verb), JJ (adjective), RB (adverb), IN (preposition), PRP (pronoun), O (Others)], apply POS tagging to the below mentioned sentences. Please use the convention '***word_POS***'.

   I. You can bank on me for money, but make sure to pay your unfinished bank installments before that. (1 mark)

   II. The round table, dance in a round, come round and see us. (1 mark)

   III. An Argentine international, Messi is his country's all-time leading goal scorer. (1 mark)

   IV. At youth level, Messi won the 2005 FIFA World Youth Championship, finishing the tournament with both the Golden Ball and Golden Shoe, and an Olympic gold medal at the 2008 Summer Olympics. (2 marks)

   I. [('You', 'PRP'), ('can', ‘O'), ('bank', 'VB'), ('on', 'IN'), ('me', 'PRP'), ('for', 'IN'), ('money', 'NN'), (',', 'O'), ('but', ‘O’), ('make', 'VB'), ('sure', 'JJ'), ('to', O'), ('pay', 'VB'), ('your', 'PRP'), ('unfinished', 'JJ'), ('bank', 'NN'), ('installments', 'NN'), ('before', 'IN'), ('that', 'O’'), ('.', 'O')]

   II. [('The', 'O'), ('round', 'NN'), ('table', 'NN'), (',', 'O'), ('dance', 'NN'), ('in', 'IN'), ('a', 'O'), ('round', 'NN'), (',', 'O'), ('come', 'VB'), ('round', 'NN'), ('and', 'O'), ('see', 'VB'), ('us', 'PRP'), ('.', 'O')]

   III. [('An', 'O'), ('Argentine', 'JJ'), ('international', 'JJ'), (',', 'O'), ('Messi', 'NNP'), ('is', 'VB'), ('his', 'PRP'), ('country', 'NN'), ("'s", 'O'), ('all-time', 'JJ'), ('leading', 'JJ'), ('goal', 'NN'), ('scorer', 'NN'), ('.', 'O')]

   IV. [('At', 'IN'), ('youth', 'JJ'), ('level', 'NN'), (',', 'O'), ('Messi', 'NN'), ('won', 'VB'), ('the', 'O'), ('2005', 'O'), ('FIFA', 'NN'), ('World', 'NN'), ('Youth', 'NN'), ('Championship', 'NN'), (',', 'O'), ('finishing', 'VB'), ('the', 'O'), ('tournament', 'NN'), ('with', 'IN'), ('both', 'O'), ('the', 'O'), ('Golden', 'NN'), ('Ball', 'NN'), ('and', 'O'), ('Golden', 'NN'), ('Shoe', 'NN'), (',', 'O'), ('and', 'O'), ('an', 'O'), ('Olympic', 'NN'), ('gold', 'NN'), ('medal', 'NN'), ('at', 'IN'), ('the', 'O'), ('2008', 'O'), ('Summer', 'NN'), ('Olympics', 'NN'), ('.', 'O')]

   (b) Devise and apply a rule (on tags) for tagging NN in above sentences. Report total number of true positives, false negatives, and false positives. You'll get 2 bonus marks for F1-score >= 75% and 1 mark for 75% > F1-score >= 40%. (Bonus: 2 marks)

   No standard answer. The evaluation will be based on the student's solution.