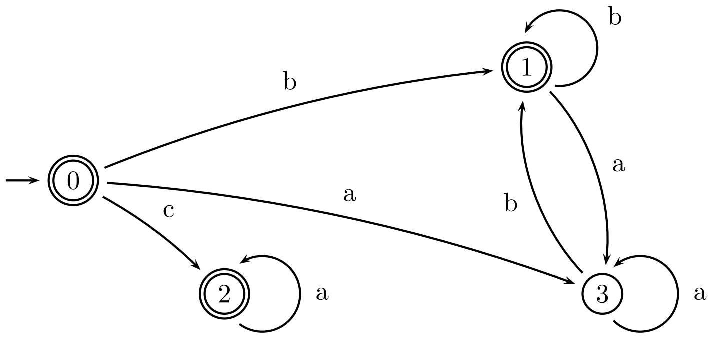
Assignment 2: Markov Models + n-grams

L645/B659, Sandra Kuebler

DUE: at beginning of class on Thursday, September 25

1. **FSA** Consider the following finite-state automaton:



   (a) Encode the FSA in terms of matrices, including initial and final states.

   (b) Test whether the following strings are accepted by the FSA, list the state sequence:
   c
   abbab
   bababa

   (c) Describe the language that is accepted by the FSA as a regular expression.

   10 pts.

2. **Markov Chains**

   There are three telephone lines, and at any given moment 0, 1, 2 or 3 of them can be busy. Once every minute we will observe how many of them are busy. This can be described as a (finite) Markov chain by assuming that the number of busy lines will depend only on the number of lines that were busy the last time we observed them, and not on the previous history.

   Use the following matrices to answer the following questions. You can use online matrix multipliers, e.g. at `http://wims.unice.fr/wims/en_tool~linear~matmult.html`. Please explain your answers.

$$P = \begin{array}{c} \\ s_0 \\ s_1 \\ s_2 \\ s_3 \end{array} \begin{array}{cccc} s_0 & s_1 & s_2 & s_3 \\ \hline \end{array}$$

$$P = \begin{array}{c|cccc} & s_0 & s_1 & s_2 & s_3 \\ s_0 & 0.2 & 0.5 & 0.2 & 0.1 \\ s_1 & 0.3 & 0.4 & 0.2 & 0.1 \\ s_2 & 0.1 & 0.3 & 0.4 & 0.2 \\ s_3 & 0.1 & 0.1 & 0.3 & 0.5 \end{array}$$

$$v = [0.5\ 0.3\ 0.2\ 0.0]$$

   (a) What is the probability that after 4 steps exactly 3 lines are busy?

   (b) What number of lines being busy has the highest probability after 4 steps?

   If you are enrolled in B659, write a program to calculate this. Submit your code and an output.

   10pts.

3. *n*-**grams**

   (a) Write a program that extracts trigrams from a text. The trigrams should be based on sentences, i.e. you have to assume sentence-beginning and -end markers. The text will have one sentence per line. Punctuation is split off. Run the *n*-gram extraction on the files `hw2-train.txt` and `hw2-test.txt`.

   (b) Calculate the trigram perplexity of text `hw2-test.txt` given `hw2-train.txt`.

   15pts.