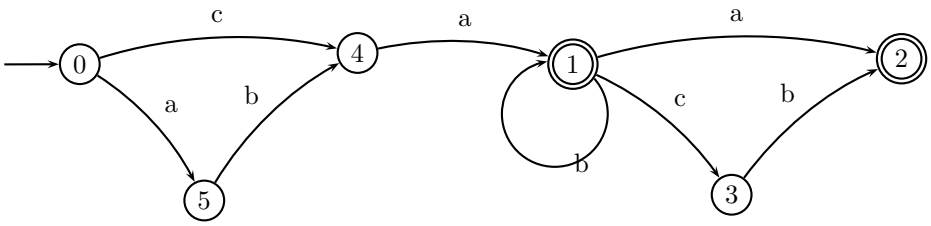
Exercise sheet 1

(Submit as a plain text email message to dm@ling.osu.edu before or on Thursday, Jan. 10)

1. Consider the following finite-state machine:



   (a) Which of the following sequences does it accept? (1) *ab* (2) *ca* (3) *cba* (4) *cabbb* (5) *ababa* (6) *bacb* (7) *cabcb* (8) *ababcbc* (9) ϵ (10) *cabbbab*

   (b) Write a regular expression which characterizes the same language as this network.

2. Go to the site `http://www.lexmasterclass.com/exercises/regex/index.html` and do the exercises 1 to 4, sending me the four regular expressions. There are multiple correct possibilities; try to find as short an expression as possible.

3. The following brute-force regular expression recognizes all English numbers from 1 to 99 in a version of the British National Corpus (BNC) sampler corpus that has been tokenized to list each word on a separate line:

```
egrep  '^(one|two|three|four|five|six|seven|eight|nine|ten|eleven|twelve|thirteen|fourteen|fifteen|sixteen|
seventeen|eighteen|nineteen|twenty|twenty-one|twenty-two|twenty-three|twenty-four|twenty-five|twenty-six|
twenty-seven|twenty-eight|twenty-nine|thirty|thirty-one|thirty-two|thirty-three|thirty-four|thirty-five|
thirty-six|thirty-seven|thirty-eight|thirty-nine|forty|forty-one|forty-two|forty-three|forty-four|
forty-five|forty-six|forty-seven|forty-eight|forty-nine|fifty|fifty-one|fifty-two|fifty-three|fifty-four|
fifty-five|fifty-six|fifty-seven|fifty-eight|fifty-nine|sixty|sixty-one|sixty-two|sixty-three|sixty-four|
sixty-five|sixty-six|sixty-seven|sixty-eight|sixty-nine|seventy|seventy-one|seventy-two|seventy-three|
seventy-four|seventy-five|seventy-six|seventy-seven|seventy-eight|seventy-nine|eighty|eighty-one|
eighty-two|eighty-three|eighty-four|eighty-five|eighty-six|eighty-seven|eighty-eight|eighty-nine|
ninety|ninety-one|ninety-two|ninety-three|ninety-four|ninety-five|ninety-six|ninety-seven|ninety-eight|
ninety-nine)$' /home/scratch/dm/bnc-samp.txt
```

   It returns 25063 matches.

   Clearly the above regular expression is missing generalizations about the structure of English numbers, which would allow it to be expressed more compactly. To improve on this, use `egrep` to test and refine the regular expression and report the regular expressions you tried and the shortest one you come up with. Note that each expression should find the same number of matches as the brute-force expression above.

   You can find the BNC sampler corpus at `/home/scratch/dm/bnc-samp.txt`. Beware, this is a large file (2,427,451 tokens), so do **not** copy it into your home directory!

   As background reading, have a look at Stuart Robinson: "Grep for Linguists", which you can find at:

   `http://arts.anu.edu.au/linguistics/misc/comp_resources/grep.html`

   An extra question to think about: Do you think each number actually occurs in this over two million word corpus of English or not? If not, how many of the 99 numbers do you expect to be missing?