Mehryar Mohri
Speech Recognition
Courant Institute of Mathematical Sciences
Homework assignment 1
Due: October 12, 2008

1. Give regular expressions describing the following languages:

   (a) The set of strings of $\{a, b\}^*$ starting with a and ending with a.

   (b) The set of strings of $\{a, b\}^*$ containing at most two consecutive a's.

   (c) The set of strings of $\{a, b\}^*$ containing exactly two occurrences of ab's.

2. Given the alphabet $\Sigma = \{0, 1, \ldots, 9\}$,

   (a) create an automaton that accepts numbers in the range $0 - 999999$.

   (b) create a transducer that maps numbers (in the range $0 - 999999$) represented as strings of digits to their English read form, e.g.,
   1 → one
   11 → eleven
   111 → one hundred eleven
   1111 → one thousand one hundred eleven
   11111 → eleven thousand one hundred eleven

   (c) Randomly generate several numbers both as strings of digits and in their read form.

   (d) Create a weighted automaton over the real semiring associating to each sequence of digits its integer value (*hint*: here is some information about the automaton: it can be constructed with just two states and the initial state has a self-loop with weight one for each digit). You should prove that the automaton is correct.

   (e) Use the weighted automaton of the previous question and the transducer previously constructed to create a weighted automaton associating to an English sequence of the type "one hundred eleven" the number in the range $0 - 999999$ that it represents. Check the correctness of the weighted automaton by applying it to "eleven thousand two hundred fifteen".

3. Given the alphabet $\Sigma = \{a, b, \ldots, z, \langle space \rangle\}$,

   (a) create a transducer that implements the *rot13* cipher – $a \rightarrow n, b \rightarrow o, \ldots, m \rightarrow z, n \rightarrow a, o \rightarrow b, \ldots, z \rightarrow m$,

   (b) encode the message `"my secret message"` (assume $\langle space \rangle \rightarrow \langle space \rangle$),

   (c) decode the encoded message from above.

4. Construct a finite-state transducer that maps any string to the set of its substrings ($y \in \Sigma^*$ is a substring of $x \in \Sigma^*$ when there exists $u, v \in \Sigma^*$ such that $x = uyv$). Create a similar weighted transducer over the real semiring and explain how it could be used to count the number of occurrences of a sequence x in a text t.