CS6320, Fall 2016
Dr. Mithun Balakrishna
Homework 1
Due September 25th, 2016 11:59pm

A. Submission Instructions:

- Submit your solutions via eLearning.
- Please submit a single zip file with the following files:
  - For programming questions:
    - Source code file(s) in C/C++, Java, or Python. For using any other programming language, please get prior approval from the TA.
    - A ReadMe file with instructions on how to compile/run the code.
  - For all other questions, a PDF/Doc/PS/Image file with the solutions.
- Late Submission Penalty:
  - up to 2 hours late — 10% deduction
  - 2 - 4 hours late — 20% deduction
  - 4 - 12 hours late — 35% deduction
  - 12 - 24 hours late — 50% deduction
  - 24 - 48 hours late — 75% deduction
  - more than 48 hours late — 100% deduction (zero credit)

B. Problems:

1. Regular Expressions (25 points)

Write regular expressions for the following. You may use either Perl/Python notation, but make sure to say which one you are using. By "word", I mean an alphabetic string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth.

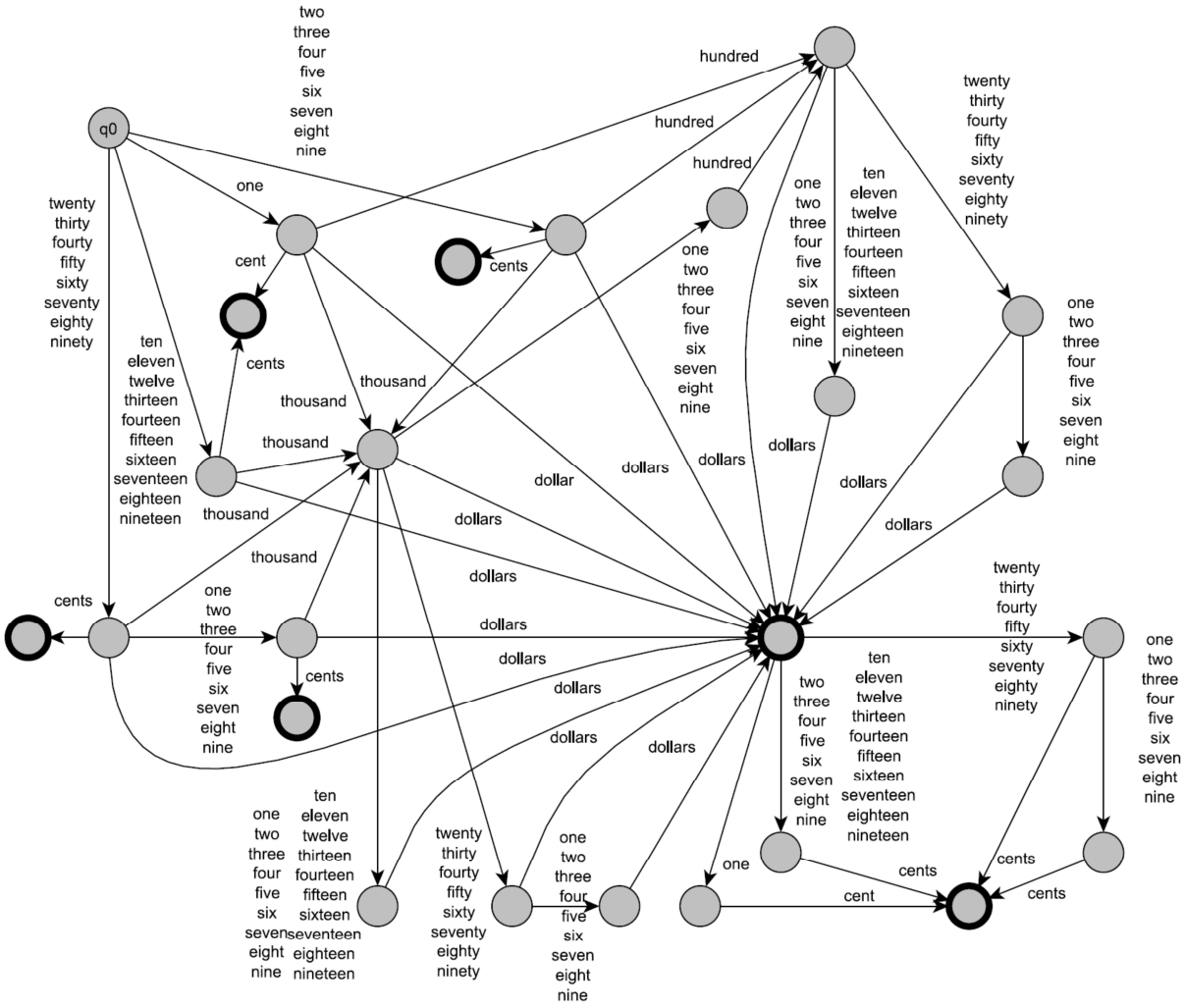
1. the set of all alphabetic strings
   `Solution: [a-zA-Z]+`
2. the set of all lower case alphabetic strings ending in a b
   `Solution: [a-z]*b`
3. the set of all strings with two consecutive repeated words (e.g., "Humbert Humbert" and "the the" but not "the bug" or "the big bug")
   `Solution: ([a-zA-Z]+)\s+\1`
4. the set of all strings from the alphabet {"a", "b"} such that each "a" is immediately preceded by and immediately followed by a "b"
   ```
   Solution:
      (b+(ab+)*)?
         OR
      b+(ab+)+
   ```
5. all strings that start at the beginning of the line with an integer and that end at the end of the line with a word
   `Solution: ^\d+\b.*\b[a-zA-Z]+$`

   Note: There can be correct regular expressions (different from the solutions listed above) to the handle the above questions.

2. Money! (25 points)

Complete the FSA for English money expressions (Slide 91 or Fig. 2.15 in text book). You should handle amounts up to $100,000, and make sure that “cent” and “dollar” have the proper plural endings when appropriate. Formulate the problem precisely, making only those distinctions necessary to ensure a valid solution. Draw a diagram of the complete state space.

Solution:



Note: There can be other correct FSAs (different from the solution listed above) to the handle the above question.

3. Bigram Probabilities (50 points):

An automatic speech recognition system has provided two written sentences as possible interpretations to a speech input.

S1: The president has relinquished his control of the company's board.

S2: The chief executive officer said the last year revenue was good.

Using the bigram language model trained on Corpus A (provided as Addendum to this homework on eLearning), find out which of the two sentences is more probable. Compute the probability of each of the two sentences under the three following scenarios:

i. Use the bigram model without smoothing.
ii. Use the bigram model with add-one smoothing
iii. Use the bigram model with Good-Turing discounting.

Write a computer program to:

A. Compute the bigram counts for any given input. Apply your program to compute the bigram counts on Corpus A.
B. For each of the three scenarios, construct the tables with the bigram counts for the two sentences above.
C. For each of the three scenarios, construct the table with the bigram probabilities for the sentences.
D. For each of the three scenarios, compute the total probabilities for each sentence S1 and S2.

Solution: Programming question.

4. Smoothed Unigram/Bigram Probabilities (25 points):

Add an option to your program from Question 3 to do Good-Turing discounting. Your program's correctness will be evaluated using the same unseen corpus as Question 3.

Solution: Programming question.