Linguistics 409 Homework 1
Regular Expressions Due: January 22, 2012

1 Programmers: (30pts)

Please do problem 2.2 from the J&M book (page 43). Have fun, but don't go nuts here. Just implement *at most* a dozen patterns. Provide (a) your working, commented source code and (b) output from an example run.

1 Non-programmers: (30 pts)

Using the NFSA in problem 2.8 (page 44), please provide the following:

1. The state-transition table
2. A regular expression
3. A few examples of grammatical sentences in this language.

2 Everyone: (40pts)

Please do problem 2.9 (page 44). Please do not implement this in a programming language, even if you can. Just use pseudo-code to describe what the changes to D-RECOGNIZE should look like. The original, buggy version of D-RECOGNIZE is given in the lecture slides and on page 29 of the textbook.

3 Everyone: UNIX & Regular Expressions (30 pts)

For each step, please provide the command you typed and, unless asked not to, the actual output it produced. Hint: the desired output, when required, will never be more than a few lines. Another hint: not all of these are most easily solved with regular expressions.

1. (2 pts) Download our corpus `http://www.gutenberg.org/cache/epub/730/pg730.txt` (no output required)
2. (2 pts) Write a regular expression to match all occurrences of 'Dodger' in the corpus. Use grep to execute this regexp and provide a count of the number of matches.
3. (2 pts) What is the most common word in this corpus? (full output not needed)
4. (2 pts) What is the least common word? (full output not needed)
5. (2 pts) How many times does 'incorrigible' appear?
6. (20 pts) In this corpus, the word 'lord' is sometimes pronounced with the final 'd' and sometimes without it. It is also sometimes capitalized and sometimes not. Write a regular expression to match these variants. Provide a count of how many times each variant occurs.