Programming in Perl, Spring 98
Project 1
Due : Wednesday, April 22 at midnight
Weight : 20% of course grade
Instructor : Louis Ziantz

This assignment consists of two parts. It should be submitted according to the *Submission Guidelines*. Don't forget to comment your program, providing a general description of the Perl script along with inline comments for important or complex portions of code.

1 Regex for Floats

For the first part, you are to develop a regular expression to recognize floating point numbers of the form described here. I would recommend doing this in a piece by piece fashion using scalar variables to store different subexpressions of the final regular expression. Store your Perl script in float.plx.

For our purposes, a floating point number consists of

1. an + or - (optional – one or neither of them can be present), followed by
2. one or more digits, followed by
3. a decimal part, followed by
4. an exponent part (optional – need not be present to match)

The decimal part consists of

1. a decimal point, followed by
2. one or more digits

The exponent part consists of

1. the letter e or E, followed by
2. an + or - (optional – one or neither of them can be present), followed by
3. one or more digits

The following are examples of floating point numbers that would be matched based on the above rules:

`6.35   -3.6   +5.4   5.0e1   2.5E+2   4.4e-10   0.5`

The following are examples of substrings that would not be matched based on the above rules:

`abc   3-e6   5+.5   6   5.   .5   5e1`

Take note of the last four. Integers should not be matched. And while the last 3 are valid in most programming languages, the rules given above exclude them to simplify things slightly.

2 Library Catalog Search

This part of the assignment involves building a simple search engine in Perl to search a library catalog. Call the script catalog.plx.

The script will obtain queries from a query file called query.input. A single query stored in the query file has the following format:

- a query tag which identifies the query
- one or more search strings each separated by a single blank
- the command AUTHOR, TITLE, or SUBJECT
- the command OR or AND
- the command SUBSTRING or EXACT

The above items are each on their own lines in the file. Some example queries:

```
1 # find books with perl OR cgi as subjects
perl cgi
SUBJECT
OR
EXACT
2 # find by author: use last or first or last,first to specify
Cormen Susan Rivest,Ronald
AUTHOR
AND
EXACT
3 # find titles with data AND structures as substrings
data structures
TITLE
AND
SUBSTRING
```

Once the five lines comprising a single query are read, the query tag should be written out on a line by itself to a query output file, call it query.output. The search strings should also be separated and stored in their own array. They will be used to create one or more regular expressions that will be employed to search through a database file named catalog.dat based on the commands read from the query file.

The format of a line in the database file is as follows:

- an author list, followed by a vertical bar (|), followed by a title, followed by a vertical bar, followed by a list of subjects that the book has information on, followed by a vertical bar, followed by the book status. A line ends with a newline.
  - The author list consists of authors' full names separated by semicolons. An author's full name is composed of the author's last name, a comma, and the author's first name.
  - The title is simply the title of the book with single spaces between words.
  - The subjects in the subject list are separated by single spaces.
  - The status is either available or due <date> where date is the day the book is due back given as <last-two-digits of year>-<two-digit-month>-<two-digit-day>.

An example line of the database file:

```
Schwartz,Randal;Christiansen,Tom|Learning Perl|perl cgi programming|due 98-04-22
```

The query commands should be handled as follows (all searches should be case insensitive):

- An AUTHOR command indicates that the search should occur on the author list portion of each line in the database. You may assume that if the user searches for a first and last name, they will enter it as `lastname,firstname` (i.e., you may search for it as a single string). The TITLE command means that the search should occur on the title portion of each line in the database, and the SUBJECT command indicates the search should be done on the list of subjects. Note that this command simply directs the target of the search; the search itself is no different in each case.
- An OR command means that if any of the strings are found in the target of the search, we have a hit. An AND command indicates that if **all** of the strings are found in the target, we have a hit. Note that all strings must be present, but not in any particular order. For the AND, you may wish to use regular expressions for each search string, rather than trying to create a single regular expression. An OR can be done with a single regular expression built with the appropriate glue.
- The command EXACT means the search string must match a word in the target of the search exactly while the SUBSTRING command indicates that it need only match a portion of a word. (Note: even when you are doing an EXACT search on a TITLE, you need only match individual words in the title exactly, **not** the title as a whole.)

For each query, the engine should search through the entire database file. If a search performed on a target yields a hit, a line of the following form should be written out to the query output file along with a newline:

```
<author list>|<title>|<status>
```

For example, query 1 above would produce the following line in query.output:

```
Schwartz,Randal;Christiansen,Tom|Learning Perl|due 98-04-22
```

Once the entire database file has been searched and all hits have been written out to the query output file, a new query should be read from the query file and the process should be repeated until the query file is exhausted.

Example query.input, catalog.dat, and query.output files will be made available on the class webpage.