Programming in Perl, Fall 97
Project 1
Due : Wednesday, November 19 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. Please get started on the assignment as soon as possible.

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.pl.

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 Search Engine

This part of the assignment involves building a simple search engine in Perl called search.pl.

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 OR or AND
- the command SENSITIVE or INSENSITIVE
- the command SUBSTRING or EXACT

The above items are each on their own lines in the file. An example of a query in the file is as follows:

```
1
perl cgi
OR
INSENSITIVE
EXACT
```

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 be create one or more regular expressions that will be employed to search through a database file named sites.dat based on the commands read from the query file. The commands should be handled as follows:

- An OR command means that if any of the strings are found in a line read from the database file, we have a subject match. An AND command means that if **all** of the strings are found in a line, we have a subject match. 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 SENSITIVE means the search should be case sensitive; INSENSITIVE means it should be a case insensitive search.
- The command EXACT means the search string must match a word in the line read from the database file exactly while the SUBSTRING command indicates that it need only match a portion of a word.

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

- a web address, followed by a colon, followed by a list of subjects that the web page has information on. The subjects are separated by single spaces and the line ends with a newline.

An example entry of the database file:

```
www.espn.com:sports baseball football basketball hockey
```

For each query, the engine should search through the entire database file. If a search performed on a line yields a subject match, the web address (and only the web address) should be written to the query output file along with a newline. Once the entire database file has been searched and all subject matches 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.