

CSE 311: Foundations of Computing I

EXTRA CREDIT: Implement Grep!

In this Extra Credit, you will implement the `grep` program using the constructions we have learned in class. The general idea is the following:

Given (1) a regular expression, and (2) a file, your program will find all the lines in the file that match the regular expression.

For instance, if your regular expression were `"bb ∪ (aa*)"`, and the file had the following lines:

```
1      aaaaaa
2      bb
3      ccaacc
4      bbbbbbbb
5      abc
6      cc
7      aaaaaaa
```

Then, your program would ultimately return lines 1, 2, 3, 4, 5 and 7. In particular, note that if *any* part of the line matches the regular expression, you should include that line.

These instructions will guide you through approaching this program in several separate phases:

- Write a CFG for our brand of regular expressions. Your program will use this CFG to parse the regular expression input!
- Read the existing parts of the NFA class to understand how it works, and then implement simulation of a string on NFAs.
- Construct NFAs for the regular expression that the user gives using the construction from lecture.
- Finally, put it all together by reading in the regex and file and outputs the right lines.

Note: You will need to type in "commandline arguments" for this assignment. To do this in JGrasp, open the Build menu and click "Run Arguments". Then, type the necessary arguments into the "Run Arguments" box at the top of the screen.

Task 1: Writing A CFG For Regular Expressions

Your first task is to write a CFG for the following grammar for regular expressions:

- `'` is a regular expression (it means the same as \emptyset).
- `-` is a regular expression (it means the same as ε).
- `a, b, ..., z, 0, 1, ..., 9` are all regular expressions.
- If A and B are regular expressions, then $A|B$ is a regular expression (this is union).
- If A and B are regular expressions, then AB is a regular expression.
- If A is a regular expressions, then A^* is a regular expression.
- If A is a regular expressions, then (A) is a regular expression.

(Note that this syntax is different from class, but it's the things with the identical meanings. We choose this alternate syntax, because it's easier to type!)

We have provided a sample grammar for *arithmetic expressions* in the `arithmetic-grammar.cfg` file; this example will be helpful as you construct a CFG for the regular expressions above. You should put

your regular expression CFG in the file `regex-grammar.cfg`. To enter the regular expression in the way the program is set up to parse it, you should be aware of the following:

- Omit arrows; they are unnecessary.
- Capital letters will be interpreted as variables; everything else will be interpreted as terminals.
- You do not need to put bars between the terms of the CFG; a space is enough to indicate that it is a choice.
- You may not use ϵ in your CFG.

You should make sure your grammar is not ambiguous. The grammar given for arithmetic expressions is set up to “encode” the precedence order for arithmetic expressions. That precedence order is (from highest to lowest): digit concatenation, parentheses, multiplication/division, addition/subtraction.

For reference, precedence for regular expressions works as follows (from highest to lowest): parentheses, star, concatenation, union. You should use a similar idea as is used in the arithmetic expression grammar for your regular expression grammar.

The Java program `Grep.java` is set up to let you test your grammar file. (Eventually, it will contain your implementation of Grep.) If you give the program a regular expression as an argument, it should tell you that it successfully parsed it (and it should fail otherwise).

Once you’re relatively sure you have it working, move on to Task 2.

Task 2: Finishing the NFA Class

Our Java classes that implement NFAs are the following: `NFA.java`, `FSMTransition.java`, `FSMState.java`.

The biggest (NFA) is the one you will be primarily editing. `FSMTransition` represents a *single* arrow in the NFA; `FSMState` represents a *single* state in the NFA. First, you should read the existing parts of the classes to familiarize yourself with them. Then, you should implement the `read(String S)` method of `NFA` which should return true if and only if the NFA accepts the string *S*.

Task 3: Regular Expression to NFA Conversion

This is the most involved of the tasks, but, luckily, we’ve given you an example of an implementation of a similar method for the arithmetic expressions in the class `ArithmeticExpressionEvaluator` (which takes in an arithmetic expression and evaluates it to a number). Your task will be to take in the parsed regular expression a user entered, and convert it (using the construction from class) to an NFA (using the `NFA` class).

To do this, you will only have to edit the `makeNFAFromRegex` method in the `Grep` class. You will have a case for each of the parts of your grammar. The `ASTNode` class is just a convenient way of saying “we have part of a parsed regular expression”. The methods of `ASTNode` let you figure out exactly what type of regular expression you have. Once you have cased on which type of regular expression the argument is, you just call the method recursively and put the results together with the constructions from class.

Task 4: Finishing Grep!

You’ve already done all of the difficult work. All that remains is to slightly modify your `Grep` class so that it takes a second argument (a file name) and runs the NFA you created in the previous part on each line of the file. Then, it should print *only the lines that the NFA matches any part of*. *Hint:* We’ve implemented a “dot” method for you which will be useful here.