This homework covers sections 3.1–3.3. It is due in class Wednesday, March 13. Hand in a hardcopy of your solutions.

Use of the RegEx101 tool (`https://regex101.com`*): #4 specifically encourages you to use RegEx101, though it isn't required. You may use RegEx101 on the other questions, but note that* **such a tool will *not* be available on exams**. *You should attempt to solve the problem and check your work for yourself before using RegEx101 — treat it as a final check rather than as a development tool.*

While you may discuss problems with other students, you should always make the first attempt on a problem yourself and **you must write up your own solutions in your own words**. *You may not collaboratively write solutions or copy a solution that one person in the group writes up.*

1. Let L and M be languages over the alphabet $\Sigma = \{a, b\}$ where

$$L = \{a, aa\} \qquad M = \{x \in \Sigma^* \mid x \text{ ends with } b\}$$

Find the following languages — write the language out as a set or give a clear English description of the language. You do not need to justify your answers, but an explanation can help you get partial credit for an incorrect answer.

a) $L \cap M$ b) $L \cup M$ c) L^3 d) L^* e) M^*

f) ML g) LM h) $\overline{M}$ i) M^R j) $M^R M$

2. For each of the following, give a clear, concise, simple English description of the language generated by the regular expression over Σ. (Simply describing the regular expression — e.g. any number of as, then a b, then any number of as — is not an acceptable answer. This language would be better described as "strings with exactly one b".) Hint: it can help to start writing out, in some methodical way, strings matching the pattern.

(a) Let $\Sigma = \{a, b\}$.

(i) bab^* (ii) $b(ab)^*$ (iii) $(a|b)^*bb(a|b)$

(iv) $a^*(b|\epsilon)a^*(b|\epsilon)a^*(b|\epsilon)a^*$

(b) Let $\Sigma = \{a, b, c\}$.

(i) $ab(a|b|c)^*ba \,|\, aba$ (ii) $((a|\epsilon)(b|c))^*(a|\epsilon)$

3. For each of the following languages, give a regular expression that generates the language. Justify your answers by explaining why the regular expression generates the strings of the language. Be careful to note the alphabet in each case, and be careful to account for all of the strings that satisfy the given condition.

   Note that this question is about regular expressions as defined in Definition 3.2 (section 3.2) in the text, not regex patterns!

   (a) $\{x \in \{a, b\}^* \mid |x| \geq 2$ and x starts and ends with the same symbol$\}$

   (b) $\{x \in \{a, b\}^* \mid x$ contains at least one a and at least one $b\}$

   (c) $\{x \in \{a, b, c\}^* \mid |x|$ is a multiple of 3$\}$

   (d) $\{x \in \{a, b, c\}^* \mid x$ doesn't start with a $b\}$

   (e) $\{x \in \{a, b, c\}^* \mid$ every a in x is immediately followed by a $b\}$

   (f) $\{x \in \{a, b, c\}^* \mid x$ contains both ab and $ba\}$

The last exercise is based on a file `data.txt` containing an (anonymized) collection of course grade data. It contains lines of the form

```
1202,MATH 100,B-,26,7/2/2019
1196,MATH 130,A,28,7/2/2019
1196,MATH 130,NC,27,8/13/2019
1196,MATH 130,C+,19,8/27/2019
```

You can find a link for the whole file on the schedule page (along with the link for this homework).

Also note that the following *are* about regex patterns, and you should use the notation discussed in class. You are encouraged to use the RegEx101 tool (`https://regex101.com`) to develop and test your answers. The whole data file is quite large, but you can copy-and-paste parts of it for testing.

4. (a) Dates in the file are given in the format `M/D/YYYY`, where `M` and `D` are the month (1-12) and day (1-31) and `YYYY` is the four-digit year. Give a search pattern and a replace pattern to convert `M/D/YYYY` dates to the format `YYYY-M-D`.

(b) The first item on each line of the file is the four-digit code for the academic term: the first digit is always 1, the second and third digits are the last two digits of the year, and the fourth digit indicates the term (2 for Spring, 4 for Summer, 6 for Fall, 8 for Winter). Give *four sets* of search and replace patterns that could be used to transform these codes to the more user-friendly format Spring 2024, Fall 2023, etc.

(If you look carefully at the file, you may notice that the year in the code doesn't match the year in the date — that's an artifact of the manipulations done to anonymize the data. It doesn't affect this task, so no need to worry about it.)

(c) Each line of the file contains five fields, separated by commas. Only the first three fields are meaningful for your application, so you would like to discard those as well as rearrange the remaining fields and add spaces after the commas. For example, the line

```
1202,MATH 100,B,24,1/15/2020
```

should become

```
MATH 100, B, 1202
```

Give a search pattern and a replace pattern to accomplish this. Be careful to match the whole line and pull out the parts that you need. Hint: `.*` can be useful here.

(d) The third field on each line contains the grade received. This includes grades such as CR, NC, VW, and so forth. Give a search pattern to match only lines with letter grades (A, B, C, D, or F, optionally followed by a `+` or `-` sign). Be careful — some of the letters can occur in other places in the file (e.g. `MATH 130` contains `A`). Hint: make use of the commas on each line!