The program takes a regex as input and transforms it to a dfa. It also takes a set of strings as input and outputs the subset of strings that matches the regex. It does so be running the dfa on each string in the set.

The transformation process consists of 4 stages. First an e-nfa is created from an ast of the regex. It is created by a traversal of the ast. At each ast node (representing an element of the regex expression) an appropriate number of states and transitions are added to the e-nfa. The book describes the basic algorithm of how to expand the e-nfa for each kind of ast node. This implementation uses a similar algorithm but it skips many empty transitions to reduce the nfa size. For instance, the union expression would produce 4 empty transitions but it produces no empty transitions in this implementation. It simply recursively passes on the two subexpressions of the union for further construction, linking the source and destination states by transitions in the subexpression.

The e-nfa can't be simulated since it is not deterministic. In the second stage, the e-nfa is transformed to a dfa. It is done be an e-closure process. A set of states for the dfa is produced that consists of the e-closures of each state in the nfa. An e-closure is the basis node together with all nodes connected to it by an empty transition. A dfa state is thus said to contain an e-nfa state if the e-closure of the e-nfa state is a subset of the dfa.

The transistions of the dfa are produced by taking each combination of an element in the alphabet and an eclosed state (this will be a source state of the transition), and calculating a destination state by gathering all states that are the destinations of the e-nfa states in the given dfa state for the given alphabet element. When each e-nfa state is eclosed in this gathering and a union of them is created, the dfa destination state is found for the chosen dfa state and alphabet element.

The starting state is found by eclosing the start state of the e-nfa and the end states are all dfa states that contain the e-nfa end state.

The resulting dfa is often too big to use in the simulation, it will therefore be minimized in the third stage. A minimized dfa is produced by first creating an equivaence table.

Write a succinct report describing your findings. For istance, how did you decide to represent your transition relation? How many states do the intermediate automata have in your test cases? Is performance affected significantly by the sizes of the two inputs? In the last step of searching for the pattern, couldn't one instead just simulate directly the #-NFA, without determinizing and minimizing it first?

The program was written in Racket. A parser was created by using the brag module language which gives tools for easily constructing a tokenizer that extracts the meaningfull segments of the input regex and for constructing a grammar that organizes the segments into an AST in the form of a Lisp S-expression. One and the same struct was used to represent each kind of automata with elements for states, transitions, start and end states and the alphabet. A transition was represented as a triple, a list of three elements (s,d,a) where s is source state, d is destination state and a the alphabet element connecting the states.

The book describes the basic algorithm of how to expand the e-nfa for each kind of ast node. This implementation uses a similar algorithm but it skips many empty transitions to reduce the nfa size. For instance, the union expression would produce 4 empty transitions but it produces no empty transitions in this implementation. It simply recursively passes on the two subexpressions of the union for further construction, linking the source and destination states by transitions in the subexpression.

One could not simulate the e-nfa directly since it is non deterministic and the program would not know which transition edge to pick or in this implementation remember which transition edges have been picked if all correct routes would be attempted. The dfa could on the other hand be simulated directly, although it would be slow in many cases.

The most intereseting test case was number 10 that in the implementation is much slower than the other test cases. It has _ states for the e-nda, _ for the dfa and _ for the miniized dfa. A way to make it much faster is to store transitions in vectors that are indexed by states. Now they are just organized flat in a list which necessitates traversal of the entire list rather than just the part represented by the state that is the source of the transition.

The reason why the test case even terminates is because the e-nfa is created more ecnomically as described above, and an alternative implementation for the equivalence table is used than the one in the book. This implementation attempts to mark as many state relations as distinguishable as possible at the outset before attempting the algorithm from the book. This mean that most states are marked as distinguishable when the algorithm from the book begins which stops most equivalence searches at the outset for the state pairs. The premarking phase is implemented by following the transitions that lead to the end states from source states that are distinguishable from the end states. By remembering the alphabet element and source state that leads to the end state, most of a row can be marked as distinguishable in the table by equivalence testing that source state on that alphabet element to every other state. This process is recursively repeated taking the adjacent source states to the end states as a new set of destination states that are distinguishable from most other states in the table.