CS421 Theory of Computation

SP16 Programming Assignment

Due Monday March 26 by 3 pm

Suppose you want to design an NFA for pattern matching of strings related to theory of computing. The goal is to retrieve documents from a collection of research papers in text mining where those phrases appear. Design an NFA that recognizes the phrases "text mining", "information retrieval", and "text data mining". This means that the language accepted by your automaton is L = {"text mining", "information retrieval", "text data mining"}. The text may or may not have these strings within " ". Your automaton must accept these strings within " " and without " ". Write a program that **simulates** the FA by computing the set of states it is in after reading each input symbol.

1. This is a two-team student assignment.
2. Programming languages you can choose from: C, C++, Java. Your programs must run on empress.
3. The input is a file containing a small collection of documents that must be considered as sets of strings, one token is a string. Remove all punctuation first and convert the documents to all lower case (normalize) before running your FA. The output is the set of documents where the phrases above are found. The name of the file is **testdoc.txt** and is posted on Cougar Courses.
4. Upload the files with 1) your source code, 2) executable/runnable code, and 3) a READ file where you describe how to compile and run your program. Name your files using your *lastname*.
5. **Make sure that you start early to meet the deadline.** No hardcopies will be given credit. No emailed files will be given credit. Programs that cannot be compiled/run on empress will not be given credit.