College of Computer & Information Science
Northeastern University
CSG714: Theory of Computation

Spring 2008
Handout 16
14 March 2008

Take-Home Midterm Exam

Due at 1:35 PM, Friday, March 21

- **Honor code:** This exam is open-notes, open-library, and open-web. However, *no collaboration of any kind is allowed.* ("Collaboration" includes, for example, discussion or exchange of material related to the problems on the exam with anyone other than the instructor.) If you have any questions or clarifications, please ask me.
- **Policy on Cheating:** Students who violate the above rules on scholastic honesty are subject to academic and disciplinary penalties. Any student caught cheating will receive an **F** (failing grade) for the course and the case may also be forwarded to the Office of Student Conduct & Conflict Resolution for further disciplinary action.
- **Presentation of solutions:** While describing a proof or a solution, you may use any of the proofs and techniques covered in class or in the text by referencing the appropriate material, without elaboration. If you use any other sources, please include citations and write the solutions in your own words.
- **A note on grading:** Your grade on any problem will be determined on the basis of the correctness of your solution and the clarity of the description. Show your work, as partial credit will be given.
- **Best of luck!**

1. (3 × 10 = 30 points) Classifying languages

Each of the following three parts gives a definition, description, or some properties of a language. In each case, tell whether the language is:

A regular

B context-free but not regular

C Turing-decidable but not context-free

D Turing-recognizable but not Turing-decidable

E not Turing-recognizable

F there is insufficient information to tell.

Justify your answers with proofs, constructions, algorithms, or examples as needed.

If, for example, you decide that a language is context-free but not regular, you must give a context-free grammar that generates the language or a pushdown automaton that accepts the language and also prove that the language is not regular.

(a) The set of pairs $\langle M, w\rangle$ such that Turing machine M, given input w, never scans any tape cell more than once.

(b) $L = \{\langle M\rangle \mid M \text{ is a Turing Machine and } M \text{ accepts all palindromes}\}$. Assume that the input alphabet for the Turing machines in L is $\{a, b\}$. Note that the language recognized by a Turing machine in L may contain non-palindromes too.

(c) $L = \{a^i b^j \mid i \neq j \text{ and } 2i \neq j\}$.

2. (10 points) The power of non-determinism

Show that for any $k \geq 1$, any k-tape nondeterministic Turing machine running in time $T(n)$ can be simulated by a 2-tape nondeterministic TM also running in time $O(T(n))$.

(*Hint:* Use one tape of the simulating machine to guess every step of the transition function in detail and use the second tape to verify the transitions on each tape of the original machine one tape at a time.)

3. (3 × 9 = 27 points) Closure properties

For a language L, the Kleene star operation yields $L^* = \cup_{k\geq 0} L^k$. That is $L^* = \{w_1 \cdot w_2 \ldots \cdot w_k : w_i \in L, k \geq 0\}$.

(a) Prove that the class P is closed under the Kleene star operation.

For an alphabet Σ, let σ denote a function from Σ to Σ (σ need not be one-to-one). For $w = a_1 a_2 \ldots a_n \in \Sigma^*$, define $\sigma(w) = \sigma(a_1)\sigma(a_2)\ldots\sigma(a_n)$. Finally, for a given language $L \subseteq \Sigma^*$, define $\sigma(L) = \{\sigma(w) : w \in L\}$. We refer to σ as an encoding.

(b) Prove that the class NP is closed under encodings. That is, show that for any encoding σ, if L is in NP, then $\sigma(L)$ is also in NP.

(c) Prove that P is closed under encoding if and only if P = NP.

4. (9 + 9 = 18 points) Linear space and reductions

(a) Give a language A that is complete for NSPACE(n) under linear-time reductions. Show that A is in SPACE(n) iff SPACE(n) = NSPACE(n).

(b) Show that if every language in SPACE(n) reduces to a given language B using logspace reductions, then every language in PSPACE also reduces to B using logspace reductions.

5. (15 points) Document clustering

Clustering of similar documents is an important technique often used in information retrieval applications. It is common to define "similarity" by means of a distance function. Let S denote a set of documents. For any two documents X and Y, let $d(X, Y)$ denote the distance between them. (We will assume that $d(X, Y)$ is a nonnegative integer.) The smaller the distance, the more similar the two documents are. One assumption often made about the distance function is that it is a metric. To be precise, assume that (i) for all X, $d(X, X) = 0$; (ii) for all X and Y, $d(X, Y) = d(Y, X)$; and (iii) for X, Y, and Z, $d(X, Z) \leq d(X, Y) + d(Y, Z)$.

Consider the Max-Cluster problem: given a set S of documents, a distance function d, and an integer r, determine the largest subset T of S such that $d(X, Y) \leq r$ for any two documents X and Y in T.

Formulate a decision version of Max-Cluster and prove that it is NP-complete. (*Hint:* You may use a reduction from Clique.)