

CS 6320.501 - Natural Language Processing – Assignment 1

Name: Shankaranarayanan Kallidaikuruchi Ramakrishnan

NetID: sxk190109

B. Problems

1. Regular Expressions

1. the set of all alphabetic strings
 $[a-zA-Z]^+$
2. the set of all alphabetic words
 $\backslash b[a-zA-z]^+\backslash b$
3. the set of all lower case alphabetic strings ending in a b
 $[a-z]^*b$
4. the set of all lower case alphabetic words ending in a b
 $\backslash b[a-z]^*b\backslash b$
5. the set of all strings from the alphabet {"a", "b"} such that each "a" is immediately preceded by and immediately followed by at least one "b"
 $b^+(ab^+)^?$
6. the set of all words from the alphabet {"a", "b"} such that each "a" is immediately preceded by and immediately followed by at least one "b"
 $\backslash bb^+a^?b^+\backslash b$
7. the set of all strings from the alphabet {"a", "b"} that form the pattern $a^n b^m$ where $(n+m)$ is even; $n \geq 0$, $m \geq 0$, and $(n+m) > 0$
 $a(aa)^*b(bb)^*|(aa)^*(bb)^+|(aa)^+(bb)^*$

2. Write a single regular expression for identifying social security numbers in text.

$\backslash b((\backslash d\{2\}[1-9]|\backslash d[1-9]\backslash d|[1-9]\backslash d\{2\})-\backslash d\{2\}-(\backslash d[1-9]\backslash d\{3}|\backslash d[1-9]\backslash d\{2}|\backslash d\{2\}[1-9]\backslash d|\backslash d\{3\}[1-9])|(\backslash d\{2\}[1-9]|\backslash d[1-9]\backslash d|[1-9]\backslash d\{2\})\backslash d\{2\}(\backslash d[1-9]\backslash d\{3}|\backslash d[1-9]\backslash d\{2}|\backslash d\{2\}[1-9]\backslash d|\backslash d\{3\}[1-9]))\backslash b$

3. Telephone Numbers

$\backslash B\backslash +(\backslash (([1-9]\backslash d|\backslash d[1-9])\backslash)-\backslash ((\backslash d\{2\}[1-9]|\backslash d[1-9]\backslash d|[1-9]\backslash d\{2\})\backslash)-\backslash ((\backslash d\{2\}[1-9]|\backslash d[1-9]\backslash d|[1-9]\backslash d\{2\})\backslash)-\backslash (\backslash d\{4\}\backslash)|([1-9]\backslash d|\backslash d[1-9])-(\backslash d\{2\}[1-9]|\backslash d[1-9]\backslash d|[1-9]\backslash d\{2\})-(\backslash d\{2\}[1-9]|\backslash d[1-9]\backslash d|[1-9]\backslash d\{2\})-\backslash d\{4\})\backslash b$

