Name:
Number:

Ceng464 Text Mining
Quiz 1

1. Write regular expressions for the following languages. (20 pts each)
a) the set of all lower case alphabetic strings ending in a b;
b) the set of all strings from the alphabet a;b such that each a is immediately preceded by and immediately followed by a b.
*a) [a-z]*b*
b) b(b+ab+)*b**
Answers similar to these are also accepted.

2. Write regular expressions for the following languages. By "word", we mean an alphabetic string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth. (20 pts each)
a) the set of all strings with two consecutive repeated words (e.g., "Humbert Humbert" and "the the" but not "the bug" or "the big bug");
b) all strings that start at the beginning of the line with an integer and that end at the end of the line with a word.
a) (.) \1*
b) ^(\d)+.(\w)+$*
Answers similar to these are also accepted.

3. Explain the terms stemming, lemmatization, and tokenization. (20 pts)
Detailed description of these terms can be found in the book and the slides.