

Matlab String Processing Exercises

Tamara Berg, Stony Brook University

Feb 7, 2012

1 Basics

- * Create a variable whose value is the string 'hello'.
- * Create a cell array containing the strings: 'hello' 'goodbye' 'hello' 'hello goodbye' 'goodbye'. Each index in your cell array should contain one of these strings. Help strfun will show a list of string processing functions. Useful functions for this exercise: cell. To get help for a particular function use the help command on the function name, e.g. 'help cell'.
- * Write code to count how many times the (exact) string 'hello' occurs in your array. Now write code to count how many times the word hello occurs in your array. Useful functions: for, find, strcmp, isempty, strfind.

2 Single Document Processing

- * Read in all of the lines from a document (http://tamaraberg.com/teaching/Spring_12/cse364/labs/spam.txt) into a cell array. Each index in your cell should contain one line from the document. Useful functions: for, fopen, fgetl.
- * Parse each line into its constituent words using the strtok function. Store all of the words in the entire document in a cell array, one word per index.
- * Create a lexicon consisting of all of the unique words in the document.

- * Create a column vector representing how many times each lexicon word occurs in the document – This is a word vector representation for the document.

3 Document Collection Processing

- * Process a collection of documents (http://tamaraberg.com/teaching/Spring_12/cse364/labs/somespams.tar.gz) to extract their words. Note list.txt gives a list of the document filenames to process.
- * Create a lexicon consisting of all of the unique words in your document collection.
- * Create a lexicon consisting of all of the unique words that appear more than 3 times in your document collection.
- * Create a matrix, where each column of the matrix is the word vector representation for a document in the collection.