# Matlab String Processing Exercises

Tamara Berg, Stony Brook University

February 13, 2013

## 1 Getting Help

* help – the most useful function in matlab. By itself shows a list of available toolboxes.

* help strfun – using help on a toolbox, shows a list of available functions in that toolbox.

* help strfind – using help on a particular function, will give information about how the function is used and often show examples of how to use the function.

## 2 Basics

* Create a variable, $A$, whose value is the string 'hello'.

* Create a cell array, $C$, containing the strings: 'hello' 'goodbye' 'hello' 'hello goodbye' 'goodbye' 'hello' 'goodhellow'. Each index in your cell array should contain one of these strings. Note, you can create an empty cell array by setting it to an empty cell, e.g. `C = [];` and you can add a string to the end of a cell array by e.g. `C{end+1} = 'hello';`

* 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.

* Write code to find the most similar string to 'goodfellow' in your array (where similarity is measured as the number of letters in common). Useful function: intersect.

* Create a variable whose value is 'hello goodby my dog fred'. Using the strtok function, tokenize this variable into its constituent words.

# 3 Document Processing

* Read in the lines of a text file, e.g. (`http://www.textfiles.com/humor/101nos.txt`), using the fgetl function. Store the lines of this file in a cell array. Useful functions: for, fopen, fgetl.

* Parse your stored lines of text into their constituent words using the strtok function. Store all of the words in the entire document in a cell array with one word per index.

* Create a lexicon consisting of all of the unique words in the document. Useful function: unique.

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