

LIS 490IDS
Fall 2015
Homework #7

Due Tuesday Nov 10, 2015 in moodle and in class

Prof. V. Stodden

Create a subdirectory in your home directory on the LIS unix system called bin. Store your executable shell scripts from this hw in that directory.

Develop shell scripts to carry out the following tasks. You should make a separate script for each question part, and think carefully about how you name each script.

Q. 1) (5 points)

- a) Take two numbers (std input) and print their sum
- b) Take a number (std input) and check whether it is even or odd
- c) Print the following table, where x is std input:

```
x * 1 = sol
x * 2 = sol
...
x * 10 = sol
```

Of course the full 10 lines should be output in the table.
As an example, if x=2 the table would look like:

```
2 * 1 = 2
2 * 2 = 4
...
2 * 10 = 20
```

Q. 2) (10 points)

a) As standard input, take a word and filename as arguments and write a script to display:

```
File <filename> contains <word> or File <filename> does not contain  
<word>
```

based on whether file actually has the given word or not. You should replace <filename> and <word> in the above messages with the actual filename and word supplied by user.

Show the results using the State of the Union addresses as the input file (you can get it from moodle). Try out the words war, budget, Cherokee, whiskey, Africa, Australia. Print an error message if the user gives fewer or greater than the number of arguments required for script to work.

b) Now modify the script (and rename it) to return the speech (identified by President and year) within which each location of the word requested was found, including all possible cases (ie. Whiskey will be found as well as whiskey). (hint: it may be useful to start question 3 first)

Q. 3) (15 points) Carry out the following exercises on the State of the Union speeches database (available in moodle). You can do this either on your laptop or using R in unix.

- (a) Use `readLines()` to read in the speeches (available as a text file in moodle) where the return value is: character vector with one element/character string per line in the file
- (b) Use regular expressions to find `***`
- (c) Use `***` to identify the date of the speech.
- (d) Use regular expressions to extract the year.
- (e) Use regular expressions to extract the month.
- (f) Use `***` to extract the name of the president State of the union speeches.
- (g) Chop the speeches up into a list there is one element for each speech. Each element is a character vector.
- (h) Eliminate apostrophes, numbers, and the phrase: (Applause.)
- (i) Make all the characters lower case.
- (j) Split the sentences up where there are blanks and punctuation to create "words".
- (k) Drop any empty words that resulted from this split.
- (l) Create a word vector for each speech.
- (m) Normalize the word vectors to get term frequencies.
- (n) Carry out some exploratory analysis of the data: Find the number of sentences, extract the long words, and the political party.