CST 334: Operating Systems

Dr. Glenn Bruns

# AWK: all word substrings

**Purpose**. What three letter sequence do you think appears most often in English words? To build your awk skills by writing a program to find character sequences that appear in English words.

**Instructions**. In this assignment you will complete an awk script substr.awk and a bash script substr-cnt.sh. Copy the tar file shown below to a directory of your own, and use the 'tar' command to extract the files.

/home/CLASSES/brunsglenn/cst334/hw/hw9/substr.tar

**Part 1**. One of the files is substr.awk. Your job is to edit this file so it will output all of the character sequences of a certain length found in a text input file. Another files is wordlist.txt, which is a text file containing about 60,000 English words, one word per line. Here are the first lines of wordlist.txt:

a

aardvark

aardwolf

ab

aba

abaca

abacist

Here's an example of the correct output of running substr.awk on wordlist.txt:

$ awk -v N=3 -f substr.awk wordlist.txt | head -7

aar

ard

rdv

dva

var

ark

aar

Important: if a character sequences appears multiple times in a single word, substr.awk should output all occurrences.

The -v N=3 part of the command sets the parameter N of substr.awk. Parameter N controls the length of the substring. Here's what happens when N is set to 4:

$ awk -v N=4 -f substr.awk wordlist.txt | head -7

aard

ardv

rdva

dvar

vark

aard

ardw

There is also a parameter M. If M=5, then substr.awk will only consider input words of 5 characters or less. By default M is 100. Here's an example where M is 5:

$ awk -v N=3 -v M=5 -f substr.awk wordlist.txt | head -7

aba

aba

bac

aca

aba

bac

ack

**Part 2**. Another of the files is substr-cnt.sh This script is supposed to give the substrings that appear in wordlist.txt along with a count for each substring, most popular substrings first. Here's an example of what it should do.

$ ./substr-cnt.sh 3 6 | head -7

229 ing

177 ter

131 ent

130 ate

128 ine

128 ver

114 per

The output shows the most common sequences of three letters among words of length 6 or less in the file wordlist.txt. Note that the output is sorted first by decreasing number, and, in the case of a tie, by increasing lexicographic order (in other words, dictionary order). Your substr-cnt.sh script should call your awk script of part 1.

Hints:

* in substr.awk, you may want to use the length() and substr() functions of awk
* in substr-cnt.sh, consider using the sort and uniq commands. Carefully check out the options for the sort and especially the uniq command.

Testing your code: The tar file also includes test scripts and a Makefile. I may modify some of the scripts for different script and slightly different input file for grading. Make sure your code runs correctly on mlc104.

**Submission**: Submit your edited substr.awk and substr-cnt.sh on iLearn.

**Grading**. You will get 5 points for each of 6 tests that your code passes. Please submit tidy code.