

# Worksheet 6

Handout 10  
CSCI 134: Fall, 2004  
12 January

---

## Hashes

---

1. Recall that the GenBank file names for the Metazoa dataset were all numbers followed by a dot followed by “gbk.” GenBank entries are usually referred to by accession numbers, however, those files are (counter-intuitively) not named with the unique accession numbers. Write a program which will associate the number of the file name (without the “.gbk” extension) with the accession number. The accession number starts with “NC\_” and can be found in several places in the GenBank file. Several are shown here:

```
LOCUS      NC_001321    16398 bp    DNA    circular    MAM      20-SEP-2002
DEFINITION Balaenoptera physalus mitochondrion, complete genome.
ACCESSION  NC_001321
VERSION    NC_001321.1  GI:5819095
```

You can extract it from the first line, or the ACCESSION line. Write out the resulting <key,value> pairs to a file called ws6.out .

2. Modify Worksheet 5 Number 4 so that it converts all the names to uppercase and then prints them out. You can start from my version if you like, it is at:

```
~stacia/ws05/ws/my_ws5_4.pl
```

3. Write a program that will “encrypt” the text (read from STDIN) by replacing all vowels (a,e,i,o,u) with a corresponding integer starting with 1 (i.e. 1,2,3,4,5). Run your program on:

```
~stacia/ws05/data/song.txt
```

4. Modify the program from the previous question to remove the blank lines from the input. Extra credit: only remove the single consecutive blank lines and change the double blank lines to single blank lines.
5. Modify the program from the previous question to replace the vowels with other characters (such as punctuation or numbers) that resemble the vowels (e.g. @ for a).