COMP[29]041 16s2 (http://www.cse.unsw.edu.au/~cs2041/16s2/)

Filters and Regexps

Software Construction (http://www.cse.unsw.edu.au/~cs2041/16s2/)

Aims

This exercise aims to give you practice with using the Unix shell for processing collections of files, and in using Unix filters for performing analyses of files.

Assessment

Submission: `give cs2041 lab02 lab02.txt`
Deadline: either during the lab, or Monday 8 August 11:59pm (midnight)
Assessment: Make sure that you are familiar with the lab assessment criteria (lab/assessment.html).

Tasks

These week's questions mostly require you to find a shell command or pipeline to solve the task.
As you go through the lab, after you determine the answer copy the pipeline and its output into a file named `lab02.txt`. Use this template (lab/lab02.txt). Note that this lab page contains hints that the template does not, so keep this page open.

After your tutor has checked your work, submit this file using the command `give cs2041 lab02 lab02.txt`.

Searching a Dictionary with `Less`

On most Unix systems you will find one or more dictionaries containing many thousands of words usually in the directories. `/usr/share/dict/` or `/usr/dict/`.
We copied one such dictionary (lab/sh/dictionary/words.txt) for this lab exercise. To save repeatedly typing its full pathname for these exercises, make a symbolic link to the file with this command:

```
$ ln -s ~cs2041/public_html/lab/sh/dictionary/words.txt
```

(this creates a link to the file in the current directory pointing to the original file, without taking up the space of a complete copy)

The preferred viewer for (potentially very long) text files is `less`, whose name is a lame pun on that of its predecessor, `more`. `less` has regular expression search capabilities.

Within `less`:

- the space bar advances a page at a time
- **b** goes back a page
- a *number* followed by `G` goes to that line (e.g. `50G` goes to line 50)
- `G` by itself goes to the end of the file
- slash (/) followed by a *regexp* finds the next instance of the *regexp* in the file (there is no terminating slash)
- question-mark (?) followed by a *regexp* finds the previous instance of the *regexp* in the file (no terminating '?')

For regular-expression searching, make sure that you understand where `less` leaves the matching line on the screen.

Access the dictionary via the command

```
$ less words.txt
```

a. How do you use `less` to find word is on line 2000 (it's at the top of the page)?

b. How do you use `less` to find the first word in the dictionary containing a 'z' `z`

c. How do you use `less` to find the last word in the dictionary that contains the substring `ooz` ?

Applying Unix Tools to a Dictionary

Write shell pipelines using `egrep`, `cat` and `wc` which answer the following questions.

Provide both the command/pipeline and its output.

a. How many total words does the file `words.txt` contain?

b. Note that some of these words are derivates of other words (e.g. "Aberdeen" and "Aberdeen's"). Maybe these shouldn't be included in the word count. How many total words, excluding those ending in "'s", are there in the dictionary?

c. How could you use `cat -n` and `egrep` to find out which word is on line 100000? (you'll need to check the exact output format of `cat -n`)

d. How could you use `cat -n` and `egrep` to print the 700th line, and no other lines? (you'll need to check the exact output format of `cat -n`)

e. How do you use `head` and `tail` to find out what word is on line 200000

Finding Unusual Words With Regular Expressions

Now consider how we might locate various "unusual" words in the dictionary using `egrep` :

a. which words contain the characters "lmn" consecutively?

b. how many words contain "zz", but do not end in apostrophe-s ('s)?

c. how many words contain four consecutive vowels?

d. which English words contain all 5 english vowels "aeiou" in that order?
Note, the word may contain more than 5 vowels but it must contain aeiou in that order. e.g. ub**a**isda**e**ucx**io**fgde**u**sc

e. how many English words contain all 5 english vowels "aeiou" in that order?

f. **Challenge** which English words contain exactly 5 vowels and the vowels are "aeiou" in that order?

g. Still on words, we want to know how many words have another 9-letter dictionary word embedded in them.
Place all the 9-letter words in the dictionary in a file called `words9.txt` . How to do this? Use a egrep pattern anchored at both ends, that matches 9 letters.

Don't include words which contain upper case characters.

Don't include words which contain any non-letter in them, We don't want to clude words such as `Bernouilli's`).

Redirect the output of the command using the > shell notation. How many of these words are there?

Now we want to use the list of words as a (huge) search pattern. `fgrep` uses a highly optimised algorithm to do this. Its usage is

```
fgrep –f WordsToFind FilesToSearch...
```

e.g.

```
fgrep –f words9.txt words.txt
```

Devise a pipeline that uses your word list (`words9.txt`) and `words.txt` to count the number of words that have a 9-letter word as a substring. For example "luminesce" is a substring of "photoluminescence".

Again we don't want to match words that contain upper-case letters or non-letters.

In other words how many 10+ letter lower case words which contain a 9 character lower-case word.

Note there are a small number of non-English words in the dictionary which have diacritical marks (http://en.wikipedia.org/wiki/Diacritic) which are encoded as a separate character in ASCII - your regular expressions don't need to handle these correctly.

egrep ing MPs

The file `/home/cs2041/public_html/lab/sh/parliament/parliament.txt` (lab/sh/parliament/parliament.txt) contains a list of the members of the Australian House of Representatives (MPs). For example:

```
$ head parliament.txt
Hon Tony Abbott: Member for Warringah, New South Wales
Hon Anthony Albanese: Member for Grayndler, New South Wales
Mr John Alexander OAM: Member for Bennelong, New South Wales
Hon Karen Andrews: Member for McPherson, Queensland
Hon Kevin Andrews: Member for Menzies, Victoria
Mr Adam Bandt: Member for Melbourne, Victoria
Ms Julia Banks: Member for Chisholm, Victoria
Hon Sharon Bird: Member for Cunningham, New South Wales
Hon Julie Bishop: Member for Curtin, Western Australia
Hon Chris Bowen: Member for McMahon, New South Wales
```

Each line contains the name of the MP followed by the electorate they represent.

a. Write an egrep command that will print all the lines in the file where the electorate begins with W. Hint the output should be:

```
Hon Tony Abbott: Member for Warringah, New South Wales
Mr Scott Buchholz: Member for Wright, Queensland
Hon Tony Burke: Member for Watson, New South Wales
Mr Nick Champion: Member for Wakefield, South Australia
Mr Peter Khalil: Member for Wills, Victoria
Mr Llew O'Brien: Member for Wide Bay, Queensland
Ms Anne Stanley: Member for Werriwa, New South Wales
Hon Dan Tehan: Member for Wannon, Victoria
Hon Malcolm Turnbull: Member for Wentworth, New South Wales
```

b. Write an egrep command that will list all the lines in the file where the MP's first name is Andrew. Hint the output should be:

```
Mr Andrew Broad: Member for Mallee, Victoria
Mr Andrew Gee: Member for Calare, New South Wales
Mr Andrew Giles: Member for Scullin, Victoria
Mr Andrew Hastie: Member for Canning, Western Australia
Mr Andrew Laming: Member for Bowman, Queensland
Hon Dr Andrew Leigh: Member for Fenner, Australian Capital Territ
Mr Andrew Wallace: Member for Fisher, Queensland
Mr Andrew Wilkie: Member for Denison, Tasmania
```

c. Write an egrep command that will print all the lines in the file where the MP's surname (last name) ends in the letter 'y'. Hint the output should be:

```
Hon Linda Burney: Member for Barton, New South Wales
Mr Pat Conroy: Member for Charlton, New South Wales
Hon Michael Danby: Member for Melbourne Ports, Victoria
Hon David Feeney: Member for Batman, Victoria
Ms Justine Keay: Member for Braddon, Tasmania
Mr Craig Kelly: Member for Hughes, New South Wales
The Hon Dr Mike Kelly AM: Member for Eden-Monaro, New South Wales
Ms Michelle Landry: Member for Capricornia, Queensland
Hon Craig Laundy: Member for Reid, New South Wales
Hon Sussan Ley: Member for Farrer, New South Wales
Mr Rowan Ramsey: Member for Grey, South Australia
Ms Anne Stanley: Member for Werriwa, New South Wales
```

d. Write an egrep command that will print all the lines in the file where the MP's name **and** the electorate ends in the letter 'y'. Hint the output should be:

```
Mr Rowan Ramsey: Member for Grey, South Australia
```

e. Write an egrep command that will print all the lines in the file where the MP's name **or** the electorate ends in the letter 'y'. Hint the output should be:

```
Hon Linda Burney: Member for Barton, New South Wales
Mr Pat Conroy: Member for Charlton, New South Wales
Mr Chris Crewther: Member for Dunkley, Victoria
Hon Michael Danby: Member for Melbourne Ports, Victoria
Mr Milton Dick: Member for Oxley, Queensland
Hon David Feeney: Member for Batman, Victoria
Hon Ed Husic: Member for Chifley, New South Wales
Mr Stephen Jones: Member for Throsby, New South Wales
Hon Bob Katter: Member for Kennedy, Queensland
Ms Justine Keay: Member for Braddon, Tasmania
Mr Craig Kelly: Member for Hughes, New South Wales
The Hon Dr Mike Kelly AM: Member for Eden-Monaro, New South Wales
Ms Michelle Landry: Member for Capricornia, Queensland
Hon Craig Laundy: Member for Reid, New South Wales
Hon Sussan Ley: Member for Farrer, New South Wales
Mr Ben Morton: Member for Tangney, Western Australia
Mr Llew O'Brien: Member for Wide Bay, Queensland
Hon Tanya Plibersek: Member for Sydney, New South Wales
Mr Rowan Ramsey: Member for Grey, South Australia
Ms Michelle Rowland: Member for Greenway, New South Wales
The Hon Tony Smith: Member for Casey, Victoria
Ms Anne Stanley: Member for Werriwa, New South Wales
Hon Wayne Swan: Member for Lilley, Queensland
Mr Trent Zimmerman: Member for North Sydney, New South Wales
```

f. Write an egrep command to print all the lines in the file where there is any part of the MP's name or the electorate name that ends in ng. Hint the output should be:

```
Mr John Alexander OAM: Member for Bennelong, New South Wales
Hon Josh Frydenberg: Member for Kooyong, Victoria
Mr Luke Gosling: Member for Solomon, Northern Territory
Mr Andrew Hastie: Member for Canning, Western Australia
Hon Michael Keenan: Member for Stirling, Western Australia
Hon Catherine King: Member for Ballarat, Victoria
Ms Madeleine King: Member for Brand, Western Australia
Mr Andrew Laming: Member for Bowman, Queensland
Hon Bill Shorten: Member for Maribyrnong, Victoria
```

g. Write an egrep command that will print all the lines in the file where the MP's surname (last name) both begins and ends with a vowel. Hint the output should be:

```
Hon Anthony Albanese: Member for Grayndler, New South Wales
```

h. Most electorate have names that are a single word, e.g. Warringah, Lyons & Grayndler. A few electorates have multiple word names, for example, Kingsford Smith. Write an egrep command that will print all the lines in the file where the electorate name contains multiple words (separated by spaces or hyphens). Hint the output should be:

```
Hon Mark Butler: Member for Port Adelaide, South Australia
Hon Michael Danby: Member for Melbourne Ports, Victoria
Hon Barnaby Joyce: Member for New England, New South Wales
The Hon Dr Mike Kelly AM: Member for Eden-Monaro, New South Wales
Mr Llew O'Brien: Member for Wide Bay, Queensland
Hon Matt Thistlethwaite: Member for Kingsford Smith, New South Wa
Mr Jason Wood: Member for La Trobe, Victoria
Mr Trent Zimmerman: Member for North Sydney, New South Wales
```

Pipelining MPs

a. Write a shell pipeline which prints the 8 Australians states & territory in order of the number of MPs they have. It should print only the names of the states/territories. It should print them one per line

Hint: check out the Unix filters cut, sort, uniq in the lecture notes.

Hint the output should be:

```
Australian Capital Territory
Northern Territory
Tasmania
South Australia
Western Australia
Queensland
Victoria
New South Wales
```

b. Challenge: The most common first name for an MP is Andrew. Write a shell pipeline which prints the 2nd most common MP first name. It should print this first name and only this first name.
Hint: check out the Unix filters cut, sort, sed, head, tail & uniq in the lecture notes.

Counting Classes

The file /home/cs2041/public_html/lab/sh/enrollments/classes.txt (lab/sh/enrollments/classes.txt) contains a list of all CSE tute/lab classes downloaded from MyUNSW.

a. How many total classes are there?

b. Write a pipeline to print how many different courses have classes? Hint: `cut` with the `-f` option will be useful here.
Hint: the output of the pipeline should be:

```
36
```

c. Write a pipeline which will print the course with the most classes (and no other courses) and how many classes are in this course? Hint: the output of the pipeline should be:

```
     29 ENGG1811
```

d. Write a pipeline that prints the most frequently-used tut room and how often it is used? Note you have to include tut-labs. In otherwords TLB as well as TUT. Hint: the output of the pipeline should be:

```
     13 Quad G041
```

e. Write a pipeline that prints the most popular time-of-day for tuts (make sure you include tut-labs) and how many tut-labs are at that time? Hint `cut` has a `-c` option. Hint: the output of the pipeline should be:

```
     25 12-13
```

f. Write a pipeline to discover which COMP courses run the most simultaneous classes of the same type? (e.g. three tutes at the same time on the same day). This one might take a little thought, but remember that this time we're looking for whether something is duplicated or not, rather than counting unique occurrences of something.

g. Challenge: write a pipeline that prints list of the course names (only) of COMP courses that run 2 or more simultaneous classes of the same type? (e.g. three tutes at the same time on the same day).
Hint this should be the output of your pipeline:

```
COMP1917
COMP1921
COMP1927
COMP2041
COMP3331
COMP6733
COMP6771
COMP9041
COMP9242
COMP9311
COMP9331
```

Challenge: Interesting Regexps

Use `egrep` to test your answers to these questions.
Try to solve these questions using the standard regular expression language described in lectures.

a. Write a regular expression for egrep that matches any line containing at least one `A` and at least one `B`. For example:

Matching	Not Matching
Andrew's favourite Band is not	George is Brillant
ABBA	Andrew
BA	B
AB	A

So to test with egrep you might do this:

```
% cat >file1 <<eof
Andrew's favourite Band is not
George is Brillant
ABBA
Andrew
AB
BA
A
B
eof
% egrep 'REGEXP' <file1
Andrew's favourite Band is not
ABBA
AB
BA
```

b. Write a regular expression for egrep that matches any line containing only the characters `A` and `B` such that all pairs of adjacent A's occur before any pairs of adjacent B's. In other words if there is pair of B's on the line , there can not be a pair of A's afterwards.

Matching	Not Matching
ABAABAABAABBBBABB	BBAA
ABBA	ABBAA
ABAAAAAAAAAABBA	ABBABABABABAA
ABABABABA	ABBBAAA
A	BBABABABABABABAA

c. Write a regular expression for egrep that matches any line containing only the characters `A` and `B` such that the number of A's is divisible by 4.

Matching	Not Matching
AAAA	AAAAA
BABABABAB	ABABBBBBBBBBBBBBBAAA
AAAABBBBAAAA	AAAABBABBAAAA
BBBAABBBBBAABBBAAAA	BBBAABBABBBAABBBAAAA

d. Write a regular expression for egrep that matches any line containing only the characters `A` and `B` such that there are exactly *n* A's followed by exactly *n* B's and no other characters.

Matching	Not Matching
AAABBB	AAABB
AB	BA
AABB	AABBB
AAAABBBB	AAAABBBBA

If you can't invent a regular expression, can you write a shell script using egrep, sed and test, if & while which performs the same task.

Finalising

You must show your solutions to your tutor and be able to explain how they work. Once your tutor has discussed your answers with you, you should submit them using `give cs2041 lab02 lab02.txt`. Whether you discuss your solutions with your tutor this week or next week, you must submit them before the above deadline.