School of Computer Science

COMP30640

Lab 8
Regular Expressions

Teaching Assistant:	Thomas Laurent
Coordinator:	Anthony Ventresque
Date:	Friday 2nd November, 2018
Total Number of Pages:	1

1 Crossword Helper

In this exercise we will manipulate the `/usr/share/dict/words` file which is the basic vocabulary used in Unix-like systems (including Linux and MacOS X). The objective is to simulate a crossword helper - similar to the little devices you find in shops or the app you can download for your smartphones.

Write a script that asks the users to input an incomplete word - mixing letters and '?' characters - and gives all the possible words taken from the file that match it (replacing '?' by 1 letter). Example:

```
$> ./crosshelper.sh abbreviat???
abbreviating
abbreviation
```

2 Text Processing

Download a book from the Gutenberg Project, for instance this one: *James Joyce's Ulysses* (maybe something shorter if you do not want to wait too long for your scripts to execute ;-)).

1. parse the text to remove the punctuation and get one word per line. Save this new version of the document in a file. You can do this by using the `tr` function and replacing punctuation, spaces and carriage returns by new lines. You might get blank lines in your new file, `tr` has an option to deal with this.
2. list the 20 most frequent words from your text
3. remove the stop words (i.e., the words with little or no real meaning); use a list of stop words from the web (e.g., one of *these lists*).
4. list the 20 most frequent words excluding stop words