

COMS3101.004 - Perl - Homework 2

April 6, 2012

1 MTA bus number extraction (5)

Write a program `mta_extract.pl` to list all mentions of MTA bus numbers in the input file. It should recognize bus numbers in all five boroughs of NYC - Manhattan, Bronx, Brooklyn, Queens and Staten Island.

The bus number contains mainly two parts - a part denoting the borough that the bus services followed by a numeric part. Express buses maybe denoted by an M after the first part or by an X. Sometimes there could be characters such as a,b,A added at the end to specify alternate routes.

The list of bus numbers can be found at the url - <http://www.mta.info/nyct/service/bus/bussch.htm>

- Look at all the bus numbers in all 5 boroughs before writing your regular expression.
- The program should list all bus numbers mentioned in the input file, one bus number per line.
- There could be more than one mentions of bus numbers per line.

Example:

<code>\$ perl mta_extract.pl passage1.txt</code>
<u>Output:</u> M104 BxM2 Bx10 Q20a ...

2 Phone number normalization (5)

Write a program `phone_no_normalize.pl` to normalize the phone numbers mentioned in the input file. A valid phone number could be mentioned with area code in parentheses - (917) 333 4444, or 917 333 4444, or 917-333-4444 or 9173334444 or even (917)-333-4444. The program should recognize any of these formats and then substitute with the standard format (917) 333 4444 and generate an output file with all phone number mentions normalized to this standard format.

- There could be more than one phone numbers per line.
- Phone numbers won't run across lines. In other words, all parts of the number will be in the same line.
- The program should print the number of phone number mentions in the passage.

Example:

```
$ perl phone_no_normalize.pl passage2.txt
```

Output:

The passage contains 34 phone number mentions.

Note: An output file `passage2_normalized.txt` should be created with the normalized passage.

3 Twitter processing (15)

Download the 'twitter_feed.txt' file. It contains an extract from twitter using the keyword 'obama'. Each line is formatted as 'name - tweet - time'. The tweet may contain any character including '-'. Write a program to

1. read each line using regular expressions and capture the name, tweet and time fields into separate variables. (3)
2. print the percentage of tweets that contain at least one url. (6)
3. list the unique hash tags (they are words starting with a #) and their number of mentions. (6)

Example:

<pre>\$ perl twitter_process.pl twitter_feed.txt</pre>
<u>Output:</u> Percentage of all tweets that contained at least one url : 99.99% Unique hash tags: #Obama : 99 #dem : 55 ...

Disclaimer: The twitter feed is an automatically extracted dataset from Twitter between 2:06am and 2:10 am on 04/06/12. The political views, wishes, and opinions expressed in the tweets are entirely tweeter's and not mine. I have tried to remove expletives, but if something has crept in, please point it out to me.