CS 517 - Natural Language Processing

Programming Assignment #1

Due date: 04/13/16 18:00 hrs

Statement

To avoid spams, people have discovered different ways to specify their emails and phone numbers in their webpages. For example, Dr. Tang has written her email id as "ftang AT cpp.edu". In this assignment, you need to write regular expressions to extract email ids and phone numbers correctly from the input dataset you have been provided. After this, you should be able to determine Dr. Manna's original email id correctly as "ftang@cpp.edu" automatically.

Code

Download pa1.zip where the basic framework for `LineProcssor.java` has been provided. You only have to implement two methods within `LineProcssor.java`:

```
HashSet<String> findEmails(String line)
```
and
```
HashSet<String> findPhoneNumbers(String line)
```

where using regular expression, you will extract all emails and phone numbers within given line and return as a set.

Format of Phone number: To simplify the assignment, all phone numbers will be US-only even though they might be written in international format in the dataset. Output format of phone number should be ###-###-####.

Code Structure

`data/` -> This directory contains some HTML pages.
`java/` -> This directory contains all the JAVA code. You only need to modify `LineProcessor.java`
`run.py` -> A tool to compile, run & test your code.

Extract and Run

Download the file pa1.zip. Extract it. Let's say you extracted into 'pa1' directory.
From terminal:

```
$cd pa1
# To run and evaluate your code
```

```
$./run.py
```

Submission

You should only modify `LineProcessor.java` and upload only `LineProcessor.java` to the Blackboard. **Any change in the filename will not be accepted and you will not be evaluated in that case**.

Evaluation

There is a held out data-set on which the instructor will run your code and compare against a rudimentary baseline extractor. If your code can find better results than the baseline, you will get 100% for this assignment. Any lesser will reduce your score linearly.