Python with Applications, PIC16
E. Ryu
Spring 2019

UCLA

Homework 2
Due 5pm, Friday, April 19, 2019

Name your file `hw2.py` and submit on CCLE. `hw2.py` should print out the answers in a reasonable format that the (human) grader can read. Comment your code adequately.

Problem 1: The file `hw2_names.txt` contains a list of 3665 US federal judges. Load the file with the following code

```
f = open("hw2_names.txt")
names = f.read()
f.close()
```

Write code that answers the following questions:

1. How many distinct lasts names do the judges have?
2. What is the most common last name?
3. How many names include 1 or more middle names?

Remark. I advise you to use regular expressions to solve this problem, although it is possible to do it without regular expressions.

Problem 2: The file `hw2_email.txt` contains emails of Enron employees, which were made public during a fraud investigation. The file `hw2_email_short.txt` is an abridged version that is easier to use for testing and debugging. Load the file with the following code

```
f = open("hw2_email.txt")
corpus = f.read()
f.close()
```

Write code that extracts all

1. email addresses,
2. phone numbers, and
3. websites

from the document.

Problem 3: The file `hw2_ucla-catalog2018-19.txt` contains the UCLA university catalog. Load the file with the following code

```
f = open("hw2_ucla-catalog2018-19.txt")
catalog = f.read()
f.close()
```

Write code that extracts all phone numbers from the entire document. Assume phone numbers are written in the format `ddd-ddd-dddd`.

The difficulty is that some phone numbers are written over two lines and that the document is written in 3 columns. For example, see the phone number `310-825-6401` in the first column of lines 26634 and 26635. There is text between `310-` of line 26634 and `825-6401` of line 26635.

Make sure your program finds most of these phone numbers.

Problem 1 clarification. Take into account hyphenated last names. For example, the last name of 'Gilberto Gierbolini-Ortiz' is 'Gierbolini-Ortiz'. Some dutch names will have the form 'James Arnold von der Heydt'. Ignore the prefixes like 'von', 'van', or 'de' and consider them as if they are middle names. Some names include suffixes like 'II' or 'III'. Treat these as if they are middle names.

Problem 2 clarification. You may wish to look up the definition of email addresses:
`https://en.wikipedia.org/wiki/Email_address#Syntax`

Different people write phone numbers in different formats. Your regular expression should catch most of them, but it is fine if you miss a few.

Use the sequence of chracters `http` or `www` (capital or lowercase) to signal the start of a website. If someone writes

```
... go to math.ucla.edu for information about ...
```

you do not have to detect such websites.

When a website is written in two lines, it can be ambiguous as to whether the following line is indeed part of the website. In the example

```
On the website https://www.math.ucla.edu/
people you can find everybody's name...

Go to https://google.com/
then search cats ...
```

it is difficult to discern whether following text is part of the website. In this case, it is fine if your program outputs `https://www.math.ucla.edu/people` and
`https://google.com/then`.

Problem 3 clarification. Some phone numbers may be nearly impossible to correctly extract using regular expressions. Consider the following example.

```
Tutorial, four hours. Limited to juniors/seniors. Individual intensive study, with scheduled meetings to be arranged between faculty member and student. Assigned reading and tangible evidence of mastery of subject matter required. May be repeated for credit. Individual contract required. P/NP or letter 310-123-1234

Biostatistics M237 and Human Genetics M207B.) Lecture, three hours; laboratory, one hour. Requisites: Biostatistics 200B, 202B (may be taken concurrently) or equivalent coursework or consent of instructor. Covers basic genetic concepts (prior knowledge of human genetics not required). Topics include 310-555-6666 methodology underlying genetic analysis of both

equations that describe fluid flow dynamics and branching, and hierarchal networks to provide survey of models for structure and flow of vascular systems. Vascular systems are nearly ubiquitous in nature, occurring across animals, plants, and other organisms. Coverage of applications to tumor growth and 310-999-4345, sleep, allometric scaling, and other phe-
```

It is fine if you do not get all phone numbers, but you should get most of them.