Methods in Computational Linguistics 1
LING 78100
Fall 2010

Homework #4 - Regular Expressions
Due at 11:50 on November 19

1. Short Problems -- Can be answered in the interpreter

A) [5] Describe the class of strings matched by the regular expression [A-Z][a-z]*
B) [10] Identify all words in nltk.corpus.words.words() which contain a sequence of 4 or more consonants.
C) [10] Write a regular expression to find every word that starts with a lowercase letter, but contains an uppercase letter in a string using re.findall(). (For example,"iPhone" or "webTV".)

2. Longer Programing Assignment -- Processing HTML

HTML presents the content of a webpage interleaved with "tags" describing how this material should be displayed by the web browser. These tags are surrounded by angled bracket characters. In this assignment you are asked to extract data or otherwise process simple html files. You do not need to know anything about HTML in order to complete this assignment.

While these are simple files, they will display correctly in a web browser.

```
<html>
<head>
<title>Page One</title>
</head>
<body>

</body>
</html>
```

The title of a page is contained between the <title> and </title> tags.

A) [10] Write a program that reads every html file and prints each title.

Text with different sizes is displayed by surrounding the text with <h1> and </h1>, <h2> and </h2>, <h3> and </h3>, and <h4> and </h4> tags. These are called header tags.

B) [15] Write a program that reads a file, replaces every header tag, h1 through h4 with h3 tags and writes a new version of the file.

C) [20] Write a program that reads every file, and generates a frequency distribution of every word in the body of the html page. Make sure to ignore any material in html tags. (HTML tags are defined by being surrounded by angled braces.)

D) [15] Links between pages are defined using the “anchor” tag, the page that is linked to is included in the “href” field of the tag. The text within the tag is the text of the link.

For example, Go to Page 1, displays a link to “page1.html” on the text “Go to Page 1”.

Write a program that identifies every link on a page.

E) [15] Write a program that processes each html file and counts the number of links that point to each html file.