

Methods in Computational Linguistics 2
Spring 2015

Homework #2 - Regular Expressions and List Comprehensions
Due by 11:59pm on Wednesday March 11

1. Short Problems — For these, simply writing a short function (def) or block of code is acceptable.

- A) [5] Construct two lists one of adjectives, and one of nouns. Using a list comprehension, construct a list which contains a **string** containing pairs of adjectives and nouns with a space between them.
- B) [5] Change your answer from part A) so that the resulting list contains only adjectives and nouns that start with the same letter. (You may need to change your lists to make sure that there are some valid combinations here.)
- C) [5] Write a list comprehension that has the same effect as the following block of code

```
>>> result = []
>>> for word in sent:
...     word_entry = (word.lower(), len(word), 'e' in word.lower())
...     result.append(word_entry)
```

- D) [5] Identify all words in `nlk.corpus.words.words()` which contain a sequence of 4 or more consonants.
- E) [5] Write a function that returns a **list** that contains every word in a **string** that starts with a lowercase letter, but contains an uppercase letter (For example, "iPhone" or "webTV"). The function should take a string as input, then separate words from the string, then return a list of words that match the requirement.
- F) [5] Use a Regular Expression to find every word in Moby Dick that ends in "ly". Put these in a frequency distribution. What are the 5 most common words that end in "ly"? (Include the code that you used to do this with your answer.)
- G) [10] Write a regular expression that matches valid US telephone numbers. Telephone numbers are 7 numbers long. Between the 3rd and 4th digit, a spacer (period, hyphen or period) can be placed e.g. 345-6789 or 3456789 or 345.6789 or 345 6789. Additionally the phone number can be preceded by a 3 digit area code. Area codes cannot start with 0 or 1. Area codes *may* be separated from the phone number by the same spacer used in the phone number e.g. 212-345-6789 is ok 212.345-6789 is not. Area codes can also be surrounded by parentheses. If they are, there should be no spacer. Also, if an area code is present, it can be preceded by the country code 1, again, with the same spacer 1.212.345.6789, or if the area code is surrounded by parentheses, a space 1 (212) 345-6789.

2. Longer Programming Assignment -- Processing HTML

In this set of questions, the requirements are described as “write a program”. A python program should be understood as a .py file that can be run from a command line by typing “python [filename].py”.

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 to the console.

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. This program should be named ‘title.py’ and called as ‘title.py inputfile’

B) [10] 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.

This program should be named ‘header.py’ and called as ‘header.py inputfile’

C) [15] Write a program that reads every file, and generates a frequency distribution of every word in the body of the html page. The program should print the N most frequent words and their frequency - one per line. Make sure to ignore any material in html tags. (HTML tags are defined by being surrounded by angled braces.)

This program should be named ‘frequency.py’ and called as ‘frequency.py inputfile N’

D) [10] 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 and prints the link text and link destination to the console. These should be separated by a tab character. This program should be named 'links.py' and called as 'links.py inputfile'

E) [15] Write a program that processes a list of html files (Hint: maybe use the glob command) and counts the number of links that point **to** each html file. The output of this program should be a list of each html file name and the number of links that point to it, one pair printed on each line.

This program should be named 'link_analysis.py' and called as 'link_analysis.py inputfile1 inputfile2, etc.' OR 'link_analysis.py inputfilepattern'