Fall Semester 2016
Iowa State University

ENGL 520 - Computational Analysis of English

Problem Set 3
Morphological Analysis and Others
(ungraded)

1. Write a program to identify proper nouns and prepositions in a input sentence, without using any language processing tools beyond regular expressions. You can assume that the user enters grammatical input.

2. Write a program to identify passive voice in sentences.

3. Write a program that identifies proper or improper use of contractions in a sentence.

4. Write a program to identify the language of a text based on writing script, using Unicode character sequences. Each script has an associated character range. So, for this program, you can pick 4,5 different writing scripts (e.g., Devnagari, Bengali, English, Cyrillic, Greek), and write a program that takes a .txt file in one of these scripts as Input, and returns the script name as output. Read a little bit about unicode character ranges, How to use them in python etc. First chapter of Dickinson et.al. (Language and Computers) is a useful reference.

5. Write a program to identify adjectives using regular expressions (obviously, every adjective usage need not be covered. Common patterns such as words ending in -ful etc can be programmed).

6. Write a program to identify capitalization errors in a paragraph of text.

7. Write a program that calculates the type-token ratio for a given piece of text (types: all unique words, tokens: all words). For the sentence "a rose is a rose is a rose", tokens:8, types: 3. So, type-token ratio is 3/8.

8. (This is a pen-and-paper problem): Do the Phonecian fun problem from NACLO website `http://clair.si.umich.edu/naclo/practice/2013G.html`

9. (This is a pen and paper problem): Do the "Curious case of Estonian" problem from NACLO website `http://www.nacloweb.org/resources/problems/2016/N2016-C.pdf`

10. Write a program that creates a fill-in-the-blank test out of a paragraph of input, by blanking out all words that are more than 8 letters long.