CSCI-UA.0480-057

Homework Number 2

Due at Midnight after the 5th Class

1. Download regexp_corpora.zip and extract the training and test corpora:
   - all-OANC.txt (your training corpus)
   - test_dollar_phone_corpus.txt (your test corpus)
2. Create 2 Programs using regular expressions to identify the following in a corpus
   - **Program 1** should identify dollar amounts
     - Cover as many cases as possible
       - including those with words like million or billion
       - include numbers and decimals
       - include dollar signs, the words “dollar”, “dollars”, “cent” and “cents.
       - include US dollars and optionally other types of dollars
       - do not include currencies that are not stated in terms of dollars and cents (e.g., ignore yen, franc, etc.)
     - The program should return each match of your regular expression into an output file, one match per line.
     - For example, if the program matched exactly 3 cases, than it would be a short file consisting of 3 lines like:
       - $500 million
       - $6.57
       - 1 dollar and 7 cents
   - **Program 2** should identify telephone numbers
     - Attempt to handle as many cases as possible: with and without area codes, different punctuation, etc.
   - **Design and test the programs** using ***all-OANC.txt***, the training corpus you downloaded and any other corpora if you choose (but not the test corpus).
   - Then **run the program** for one last time on the test corpus you downloaded: ***test_dollar_phone_corpus.txt***
     - These are the results you should submit for grading (see below).

        - You should not use this corpus to develop your system -- you should only run on the test corpus when you are done writing the program
    - **Programming language**: the program can be in any standard programming language
3. Submit your program and the output to Gradescope as your answer for homework 2. You should submit a total of 4 files, combined into a single zip file which you should name a follows: NetID-HW2.zip, e.g., alm4-HW2.zip. . The four files should have the following names and contain the following information:
    - dollar_program.filetype -- your program file for finding dollar amounts
        - Example: dollar_program.py
        - This is your program file. It should be possible to call your program on the command line with a text file as a parameter and output regexp matches in the format indicated below. For example,

          dollar_program.py ***test_dollar_phone_corpus.txt***

          should cause some matches (dollar amounts) to be returned to the terminal.

        - The filetype should depend on the programming language -- the filetype can be .py, .sh, etc.
        - **The regexp used in your code should be easy for a grader to identify.** The code could directly include a regexp that is easy for a reader to look at or it could be made up of a collection of variables concatenated together. If you do the latter, it should be easy for the regexp to be printed out, e.g., you could have a program option that allows this.
    - telephone_regexp.filetype -- your program for finding telephone numbers. The details are basically the same as for the dollar regexp program file
    - dollar_output.txt -- this should contain the dollar amounts recognized by your program, one per line. The parts of the lines that are not part of the dollar amount should not be printed at all. 3 lines of example output might be something like this:
        - $5 million
        - $5.00
        - five hundred dollars
    - telephone_output.txt – the output file for telephone numbers, e.g.,
        - 212-345-1234
        - 777-1000
4. Programs will be graded by how well they do on the test_dollar_phone_corpus according to 2 metrics:
    - Precision: Number of Correct Answers / Number of Answers

        - If there are many answers to grade, we may use sampling to estimate precision
    - Coverage: Number of Correct Answers

Accessibility