NYUAD Course CS-AD 220 – Spring 2016
Natural Language Processing

Assignment #1
Unix Tools and Regular Expressions
Assigned Feb 4, 2016

Due Feb 18, 2016 (11:59pm)

I. Grading & Submission

This assignment is about the use of regular expressions (regex) and a set of Unix tools for quick text processing. The assignment accounts for 10% of the full grade. Section III below has a set of questions. The student needs to answer them all. The specific number of points for each question is provided. The student should submit a **PDF** file containing the answers to each question and sub-question in order. The student should also include the commands and the result of applying the commands by copying and pasting from the terminal. Each student must work alone. This is not a group effort.

The assignment is due on Feb 18 before midnight (11:59pm). For late submissions, 10% will be deducted from the homework grade for any portion of each late day. The student should upload the answer to NYU Classes (Assignment #1).

II. Before Starting

A. The United Nations Corpus

In this assignment, you will make use of the United Nations (UNCorpus), a corpus on the UN general assembly resolutions. The UNCorpus is a six-language parallel text in Arabic, Chinese, English, French, Russian and Spanish. The following paper describes the corpus:

> Alexandre Rafalovitch and Robert Dale. 2009. United Nations General Assembly Resolutions: A Six-Language Parallel Corpus. In Proceedings of the MT Summit XII, pages 292-299, Ottawa, Canada.
> URL: http://www.uncorpora.org/Rafalovitch_Dale_MT_Summit_2009.pdf

You can download the text of the UNCorpus from
http://www.uncorpora.org/files/uncorpora_plain_20090831.zip

Unzipping the file produces a text file named **uncorpora_plain_20090831.tmx**. This file will be referred to as the UNCorpus in the rest of this document.

B. Unix Tools

Revise the usage of the following Unix commands (and some of their specific options), which you will need in this assignment: *cat, wc, sort (sort –nr), uniq (uniq –c), grep (grep –e; grep –a), comm,* and *more* (the command that is). You can use the *man* command to check the usage from any Unix terminal (eg *man cat*). You can also check this online man page: http://man7.org/linux/man-pages/. Other Unix commands you may want to consider checking are: *less, tr and sed*.

Additionally, revise the use of the pipeline and I/O redirections (| and >, specifically). For a quick introduction, see http://www.westwind.com/reference/os-x/commandline/pipes.html.

Finally, we recommend using the PERL interpreter in a Unix command pipeline mode to apply regex substitutions: perl –pe '<substitute-regex>;'. It is more powerful than sed or tr commands.

C. Regular Expressions

Revise the regular expression definitions in Chapter 2 in J+M Book. There is a cheat sheet in the inside cover of the book. Here is another link to a different cheat sheet also: http://web.mit.edu/hackl/www/lab/turkshop/slides/regex-cheatsheet.pdf Another useful resource to play with is http://www.regexr.com/.

III. Questions & Answers

In the following pages you will find a number of questions and sub-questions. Please include the answers in the PDF file. All answers should be colored RED. Keep the same number of pages. If an answer needs more space, reduce the font of the answer. See Example Q0 next.

Q0: (0pt) Ignoring letter case, how many lines of text in the UNCorpus mention the term *Human Rights*?

Answer: 5,664 lines of text.

```
cat uncorpora_plain_20090831.tmx |egrep -i 'human rights'|wc
    5664  279696 1969627
```

Q1: (5pt) Describe the XML structure of the UNCorpus file.

Q2: (15 pt) The Full UNCorpus

Answer the following questions using Unix commands and regex only. Each question should be answered with one command line (possibly consisting of multiple piped Unix commands)

a. How many lines does the UNCorpus file have?

b. How many segments <seg>?

c. How many non-segments? As in tags that are not <seg> like <tuv>?

d. How many English segments does the text have?

e. How many segments exist for each languages (Chinese, Arabic,...)? (again, done in one command)

Q3: (30pt) The English UNCorpus

Answer the following questions using Unix commands and regex only. Each question should be answered with one command line (possibly consisting of multiple piped Unix commands)

a. Extract the text without XML for only the English segments and put in a file called "uncorpus.eng.txt" (Hint, use "grep –a1"). The rest of the questions are about this file. How would you verify that you did not miss any lines?

b. Count the total number of words (tokens).

c. Count the total number of unique words (types).

d. Count the total number of unique words ignoring capitalization.

e. Count the total number of pure digits tokens.

f. Count the total number of digits with non-word characters with them (e.g. 8,000.00).

g. Count the total number of words starting with capital letters.

h. What are the top 15 most common first words of sentences?

i. What are the top most common capitalized words (that are not sentence initial).

j. Count all occurrences of Roman numerals.

Q4 (15 pt): Capitalization Exercise

This question uses the file "uncorpus.eng.txt" as the corpus.

a. Extract the list of unique segment-initial capitalized words and put in file cap-1.txt. (make sure to skip sentences that start with non-letter characters.)

b. Extract the list of unique non-segment-initial capitalized words and put in file cap-not1.txt.

c. Count the number of words in both files.

d. In one command get the number of overlapping words that are in both cap-1.txt and cap-not1.txt, words only in cap-1.txt, and words only in cap-not1.txt. (Hint use comm)

e. Repeat (d) but without *comm*.

Q5: (15 pt) Histogram

This question uses the file "uncorpus.eng.txt" as the corpus.

a. Construct a histogram of the words in the corpus -- a list of words sorted by frequency (with specified frequency). Identify the 10 most frequent words (top 10) and 10 of the least frequent words (bottom 10) in the full corpus.

b. Construct a histogram based on the **first** 10,000 segments and identify top and bottom 10 words.

c. Construct a histogram based on the **last** 10,000 segments and identify top and bottom 10 words

d. Construct a histogram based on the **first** 35,000 segments and identify top and bottom 10 words

e. Compare the four sets of top 10; and the four sets of bottom 10. What words are similar, or different? Are you surprised (or not surprised) by the results? (why?)

Q6: (20 pt) Back to the Original Corpus

a. Get the top 20 (most frequent) words in English, Arabic, Spanish and Russian of the UNCorpus. You will need four separate commands. Show the lists of words in your answer.

   For this task, consider a word to simply be white-space separated (i.e. keep all punctuation and digits and separate on white space).

b. Use Google Translate to compare the meanings of these words. What words are similar, what are different? Show your work including the results of Google Translate.