Natural Language Processing

Lab #1
Unix Tools and Regular Expressions

I. Instructions:

This Lab is about the use of regular expressions (regex) and a set of Unix tools for quick text processing. Section III below has a set of questions. You should include the commands and the result of applying the commands by copying and pasting from the terminal.

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

Alternate links:

1. https://conferences.unite.un.org/uncorpus/en/downloadoverview
2. https://www.kaggle.com/boulaalamcfk/un-v10-corpus/version/1

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. If you use the Unix cluster, the corpus is available to you in

```
/home/share/nlp/uncorpora_plain_20090831.tmx
```

B. Accessing the Unix

Access the Ubuntu 18.04 terminal.

Make sure you are in your home directory.

```
cd ~
```

Create a new directory in your home directory called `NLP`

```
mkdir NLP
```

Change directory to the new `NLP` directory

```
cd NLP
```

Save your files in this new directory or download them using
wget http://www.uncorpora.org/files/uncorpora_plain_20090831.zip

Run commands on the UNCorpus file by using the absolute path (bolded below).

```
cat /home/share/nlp/uncorpora_plain_20090831.tmx
```

This file (uncorpora_plain_20090831.tmx) will be referred to as the UNCorpus in the rest of this document.

As a simple example, execute the following command to copy the first 10 lines of the corpus to **top-10.txt** in your new `NLP` directory. This is not necessary, but illustrates how you can use redirection and execute commands on the UNCorpus file.

```
head /home/share/nlp/uncorpora_plain_20090831.tmx > top-10.txt
```

C. 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://unixhelp.ed.ac.uk/alphabetical/. 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>;' or python -c <stuff> . It is much more powerful than sed or tr commands.

D. 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

III. Questions

Q1: 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)

Q2: 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.

Q3: Back to the Original Corpus

a. Get the top 20 words in English, Arabic, Spanish and Russian of the UNCorpus. (You will need four separate commands)
b. Use Google translate to compare the meanings of these words. What words are similar, what are different?