# Natural Language Processing

## Lab #1 and #2
## Unix Tools and Regular Expressions

The data is uploaded on Canvas.

------

New Link

[https://web.archive.org/web/20180831123202/http://www.uncorpora.org/files/uncorpora_plain_20090831.zip](https://web.archive.org/web/20180831123202/http://www.uncorpora.org/files/uncorpora_plain_20090831.zip)

Plain TM version (40.9/155.6 MBytes). In this version, voting segments are removed, footnotes are removed completely and symbols and lead markers are removed (but the content is kept). This is a version suitable for import into commercial TM tools, which may not be implementing full TMX spec.

United Nations General Assembly Resolutions: A Six-Language Parallel Corpus
http://www.uncorpora.org/Rafalovitch_Dale_MT_Summit_2009.pdf

Alexandre Rafalovitch, 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, August.

1) use Perl

```
cat uncorpora_plain_20090831.tmx |perl -pe 'tr/[A-Z]/[a-z]/;'|more
```

2) how many lines does the UNCorpus file has?
how many segmenets <seg>?

```
grep '<seg>' uncorpora_plain_20090831.tmx | wc -l
```

how many non <seg>

```
grep '<*>' uncorpora_20090831-sample-a.tmx | grep -v '<seg>' uncorpora_20090831-sample-a.tmx | wc -l
```

what percentage of the the file size is text vs xml?

```
cat uncorpora_plain_20090831.tmx |perl -pe 's/<seg>.*<\/seg>/<seg><\/seg>/;'|wc
1501316 2062229 30154494
```

3) How many English segments does the text have?

```
cat uncorpora_plain_20090831.tmx |grep "xml:lang=\"EN\"" |wc -l
  72339
```

4) count the segments for all the languages (Chinese, Arabic,...) using ONE command.

```
cat uncorpora_plain_20090831.tmx |grep "xml:lang=\"..\"" |sort |uniq -c|sort -nr
72339     <tuv xml:lang="ZH">
72339     <tuv xml:lang="RU">
72339     <tuv xml:lang="FR">
72339     <tuv xml:lang="ES">
72339     <tuv xml:lang="EN">
72339     <tuv xml:lang="AR">
```

5)

```
ADUAE06419LP-MX:Assignment-1 nh48$ cat uncorpora_plain_20090831.tmx |grep
"\band\b"|wc
  49036 2327159 16607612
ADUAE06419LP-MX:Assignment-1 nh48$ cat uncorpora_plain_20090831.tmx |grep
"and"|wc
  86480 4456732 31323758
```

```
grep -a1 "lang=\"EN\"" uncorpora_plain_20090831.tmx |grep "<seg>"
```

```
grep -a1 "lang=\"EN\"" uncorpora_plain_20090831.tmx |grep "<seg>" |perl -pe
's/\s*<\/?seg>//g;'|wc
  72339 2685545 18008957
```

How do you verify that you did not loose any lines?

cut all words -> one per line:

```
RESOLUTION
55/100
Adopted
at
the
81st
plenary
meeting,
on
4
December
2000,
on
the
recommendation
of
the
Committee
(A/55/602/Add.2
and
```

http://en.wikipedia.org/wiki/ASCII

```
ADUAE06419LP-MX:Assignment-1 nh48$ cat eng |perl -pe 's/ /\n/g;'|grep -v "[0-
z]"|sort|uniq -c |sort -nr
 114 •
  68 -
  36 ",
  13
   7 ".
   6 ...
   6 *
   4 )
   3 ,
   2 …
   1 ...,
   1 "
ADUAE06419LP-MX:Assignment-1 nh48$ cat eng |perl -pe 's/ /\n/g;'|grep -v "[!-
~]"|sort|uniq -c |sort -nr
 114 •
  13
   2 …

ADUAE06419LP-MX:Assignment-1 nh48$ cat eng |perl -pe 's/ /\n/g;'|grep -v "[A-Za-z]"|wc
  107474  107461  482549
```

How many word have repeated ss

```
ADUAE06419LP-MX:Assignment-1 nh48$ cat eng |perl -pe 's/ /\n/g;'|egrep "ss"|wc
   61894
```

repeated char

```
cat eng |perl -pe 's/ /\n/g;'|egrep "(.)\1"|wc
  307567
```

how many triples?

how many are digits, roman numerals, other?

```
cat eng |perl -pe 's/ /\n/g;'|egrep "(.)\1\1"|egrep  "[iIxXvVcCmMLl]"|wc
    877
```

create two files - one containign the top 10,000 lines; and another lowest 10,000 lines