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?

```
 grep -A1 '<tuv xml:lang="EN">' uncorpora_plain_20090831.tmx | perl -pe 's/<seg>//'| perl -pe 's/<\/seg>//' | perl -pe 's/<tuv xml:lang="EN">//'| perl -pe 's/--//' | grep '[^[:blank:]]' > uncorpus.eng.txt
 ###You could check total number of lines and compare to number of english lines in corpus from question 8###
```

In the above code I assume that the -- should also be removed as they follow every line.

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

```
wc -w uncorpus.eng.txt
###2685545 uncorpus.eng.txt###
```

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

```
cat uncorpus.eng.txt | perl -pe 's/\s/\n/g;' | sort | uniq | wc -w
###37033###
```

d) Count the total number of unique words ignoring capitalization

```
cat uncorpus.eng.txt | perl -pe 'tr/A-Z/a-z/;' | perl -pe 's/\s/\n/g;' | sort | uniq |wc -w
###33365###
```

e) Count the total number of pure digits tokens.

```
egrep -o '\b[0-9]+\b' uncorpus.eng.txt | wc -l
###133130###
```

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

```
egrep -o '[0-9]+([\.,][0-9]+)+' uncorpus.eng.txt | wc -l
###1721###
```

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

```
egrep -o '[A-Z][a-z]*' uncorpus.eng.txt | wc -l
###494428###
```

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

I assume sentences can only end with the following: .?!

```
egrep -o '[\.\!\?]\s[A-Z][a-z]*' uncorpus.eng.txt | perl -pe 's/[\.\!\?]\s//g;' | sort | uniq -c | sort -nr | head -15
###3703 Requests###
###2415 Calls###
###2380 Also###
###2028 Welcomes###
###1941 Decides###
###1688 Urges###
###1632 Notes###
```

```
###1607 Encourages###
###1482 Takes###
###1458 Reaffirms###
###1409 Invites###
###1044 Stresses###
### 890 Recognizes###
### 861 Expresses###
### 737 Emphasizes###
```

i) What are the top most common capitalized words (that are not sentence initial). (ASSUMING YOU WANT THE TOP 15)

```
egrep -o '[^\.\!\?]\s[A-Z][a-z]*' uncorpus.eng.txt | perl -pe 's/^[^\.\!\?]\s//g;' | sort | uniq -c | sort -nr | head -15
###19709 United###
###10167 States###
###9302 December###
###8947 Secretary###
###6228 General###
###5297 International###
###4957 Convention###
###4823 Recalling###
###4582 Committee###
###3723 The###
###3514 Member###
###3178 Commission###
###2908 Human###
###2868 Assembly###
###2569 Organization###
```

j) Count all occurrences of Roman numerals

I assume all roman numerals will be capitalized

```
egrep -o '\b[MCLDXVI]+\b' uncorpus.eng.txt | wc -l
###4778###
```