Assignment1

Mengxi Nie G48739076 eva_happylife@gwu.edu

Q1:

- **a)** /[13579][02468]/
- **b)** /[A-Za-z]\W[0-9]/
- **c)** /\b^[A-Z][a-zA-Z]*[:punct:]\b/
- **d)** /(?!)ping/
- **e)** /\b[0-9]{1,2}\.[0-9]{1,2}\.[0-9]{2}\b/

Q2:

The regular expression is

/^((25[0-5]|2[0-4]\d|((1\d{2})|([1-9]?\d)))\.){3}(25[0-5]|2[0-4]\d|((1\d{2})|([1-9]?\d)))/

I use Perl and the file is Q2.pl

OS: ubuntu14 64bit

Running Result:

```
mengxi@ubuntu:~/Desktop/perl/HW1$ chmod 755 Q2.pl
mengxi@ubuntu:~/Desktop/perl/HW1$ ./Q2.pl
Please enter an IP address
78.36.25.98
Yes, it is an IP address!
mengxi@ubuntu:~/Desktop/perl/HW1$ ./Q2.pl
Please enter an IP address
888.99.25.36
No, it isn't an IP address!
mengxi@ubuntu:~/Desktop/perl/HW1$ ./Q2.pl
Please enter an IP address
52,56.25.89
No, it isn't an IP address!
mengxi@ubuntu:~/Desktop/perl/HW1$
```

Q3:

I use Perl and the file is Q2.pl

OS: ubuntu14 64bit

Running Result:

```
mengxi@ubuntu:~/Desktop/perl/HW1$ chmod 755 Q3.pl
mengxi@ubuntu:~/Desktop/perl/HW1$ ./Q3.pl
Please enter a number:
you
It isn't a number.
mengxi@ubuntu:~/Desktop/perl/HW1$ chmod 755 Q3.pl
mengxi@ubuntu:~/Desktop/perl/HW1$ ./Q3.pl
Please enter a number:
98
The sum is 17.
mengxi@ubuntu:~/Desktop/perl/HW1$
```

Q4:

I use Python2.7 and the file is *hwk1_p4.py*. The original sentences are stored in *hwk1_p4.txt and* the result is stored in the file *hwk1_p4_result.txt.*

OS: Windows10 64bit

Q5 :

I use Python2.7 and the file is *hwk1_p5.py.*

OS: Windows10 64bit

Running Result:

```
C:\Python27\nlp>hwk1_p5.py
4 9 2 7 2 0 2 0 13 23 13
0 2
2 3
4 1
7 1
9 1
13 2
23 1
```

Q6 :

I use Python2.7 and the file is *hwk1_p5.py.* The result is stored in the file *hwk1_p6_result.txt.*

OS: Windows10 64bit

Q7 :

grep -A 1 'lang="EN"' uncorpora_plain_20090831.tmx | grep -i "Human Rights" | wc -l

Ans: 5663

Q8:

(a) *wc -l uncorpora_plain_20090831.tmx*
Ans: 1501316

(b) *grep -o '<seg>' uncorpora_plain_20090831.tmx | wc -l*
Ans: 434034

(c) *grep "<[^(s|\/)].*>" uncorpora_plain_20090831.tmx | wc -l*
Ans: 560895

(d) *grep -A 1 'lang="EN"' uncorpora_plain_20090831.tmx | grep '<seg>' | wc -l*
Ans: 72339

(e) *cat uncorpora_plain_20090831.tmx | grep -o 'lang="\w\w"' | sort | uniq -c*
72339 lang="AR"
72340 lang="EN"
72339 lang="ES"
72339 lang="FR"
72339 lang="RU"
72339 lang="ZH"

Q9:

(a) *grep -A1 'lang="EN"' uncorpora_plain_20090831.tmx | awk -F'>|<' '{print $3 > "uncorpus.eng.txt" }'*

(b) *egrep -o "[a-zA-Z-]+|([A-Z]\.)+|('[a-z]\b)" uncorpus.eng.txt | wc -w*
ans : 2583849

(c) *cat uncorpus.eng.txt | egrep -o "[a-zA-Z-]+|([A-Z]\.)+|('[a-z]\b)" | sort | uniq | wc -l*
ans: 15362

(d) *cat uncorpus.eng.txt | egrep -o "[a-zA-Z-]+|([A-Z]\.)+|('[a-z]\b)" | sort | uniq -i | wc -l*
ans:12605

(e) *grep -P -o '\b\d+\b' uncorpus.eng.txt | wc -l*
ans: 133130

(f) *uncorpus.eng.txt | egrep -o "([0-9]{1,3}(\,[0-9]{3})*(\.[0-9]+)?)|([0-9]+)" | wc -l*
ans: 133029

(g) *egrep -o '\b[A-Z]([a-zA-Z-]*|(\.([A-Z]\.)*))\b' | wc -l*
ans: 451347

(h) grep -o -P '[a-zA-Z]+.' uncorpus.eng.txt | grep -o -P '(?<=^)[a-zA-Z]+' | sort | uniq -c | sort -nr | head -15*

ans:

4821 Recalling
3713 Requests
2811 The
2524 Also
2417 Calls
2166 Adopted
2090 Noting
2054 RESOLUTION
2034 Welcomes
1969 Decides
1694 Urges
1679 Recognizing
1669 Reaffirming
1646 Notes
1606 Encourages

(i) grep -o -P '[a-zA-Z]+.' uncorpus.eng.txt | grep -o -P '(?<!^)\b[A-Z]([a-zA-Z-]*|(\.([A-Z]\.))*)\b' | sort | uniq -c | sort -nr | head -15*

ans:

19938 United
19162 Nations
14934 States
9302 December
8961 Secretary-General
8621 Assembly
8503 General
7001 Committee
5731 Convention
5595 International
3902 Declaration
3867 Commission
3726 Conference
3688 Rights
3593 Development

(j) grep -P -o '(?<!\/)\b(M{0,4}(CM|CD|D?C{0,3})(XC|X?L|L?X{1,3})(IX|IV|V?I{0,3})|M{0,4}(CM|CD|D?C{0,3})(XC|XL|L?X{0,3})(IX|I?V|V?I{1,3}))\b' uncorpus.eng.txt | wc -l

ans: 3611

Q10:

egrep -o '\b[a-zA-Z]+\b' uncorpus.eng.txt | sort | uniq -c | sort -nr | head -40

```
mengxi@ubuntu:~/Desktop/perl/HW1$ egrep -o '\b[a-zA-Z]+\b' unco
 268153 the
 176014 of
 138294 and
  99762 to
  66988 in
  36024 on
  32461 for
  24039 that
  21550 a
  21181 its
  20415 with
  20127 United
  20024 as
  19188 Nations
  17986 by
  17726 General
  15118 States
  13981 at
  13078 all
  12220 international
  11173 their
   9302 December
   9152 including
   9087 Secretary
   9084 or
   8918 development
   8621 Assembly
   7622 resolution
   7367 report
   7013 Committee
   6961 other
   6862 countries
   6832 rights
   6497 session
   6388 be
   6379 implementation
   6377 human
   6218 organizations
   6076 which
   6000 from
```

egrep -o '\b[a-zA-Z]+\b' uncorpus.eng.txt | sort | uniq -c | sort -n | head -40

```
mengxi@ubuntu:~/Desktop/perl/HW1$ egrep -o '\b[a-zA-Z]+\b' uncorpus.en
      1 abandon
      1 abatement
      1 abductees
      1 Abdullah
      1 aberrations
      1 abided
      1 abiotic
      1 abject
      1 Abkhazian
      1 absorbing
      1 absorptions
      1 abstain
      1 Abstaining
      1 abstinence
      1 Abusive
      1 Accelerating
      1 accentuated
      1 Accentuates
      1 acceptability
      1 acceptably
      1 Accepting
      1 accessed
      1 Accessible
      1 accommodated
      1 accommodations
      1 accountants
      1 accounted
      1 accredit
      1 Accreditation
      1 accrued
      1 accumulate
      1 Achieve
      1 achieves
      1 Acknowledge
      1 Acknowledgement
      1 acquainted
      1 acquiesced
      1 Acquired
      1 acres
      1 acronym
```

Top words are mostly preposition and words that are associated with the essay theme. Most bottom words come from the same root. I am not surprised by the top words result but shocked with the bottom words result because I thought the bottom words were some vocabularies which

expressed the contrary of the same. For example, an essay is about happy then some negative words appear rarely.

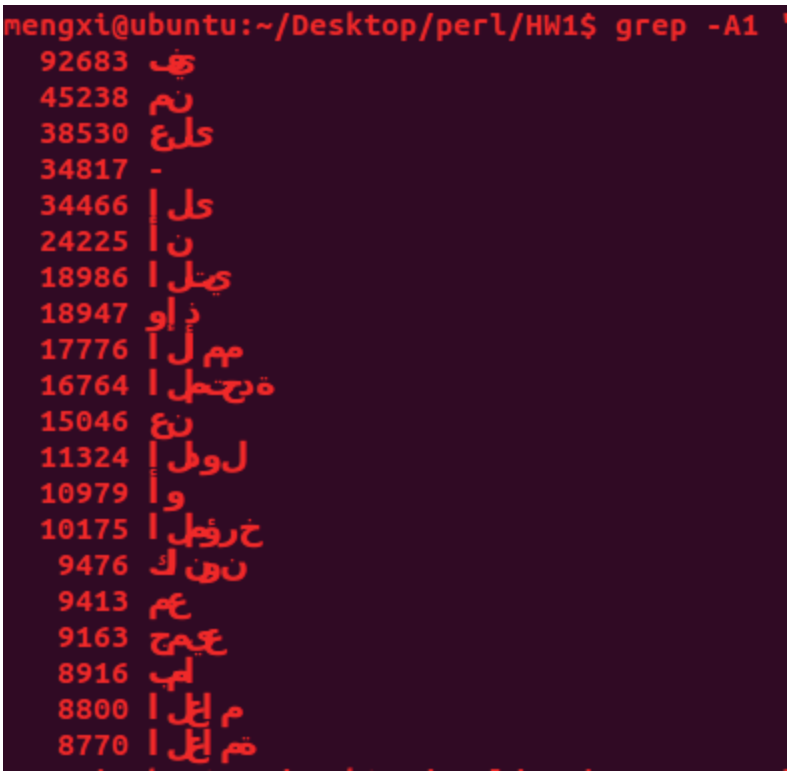
Q11:

A:

a) *grep -A1 'lang="EN"' uncorpora_plain_20090831.tmx | grep -o -P '(?<=<seg>).*(?=</seg>)'| grep -o -P '(?<=\s|\A)[^\s]+(?=\s|\Z)' | sort | uniq -c | sort -nr | head -20*

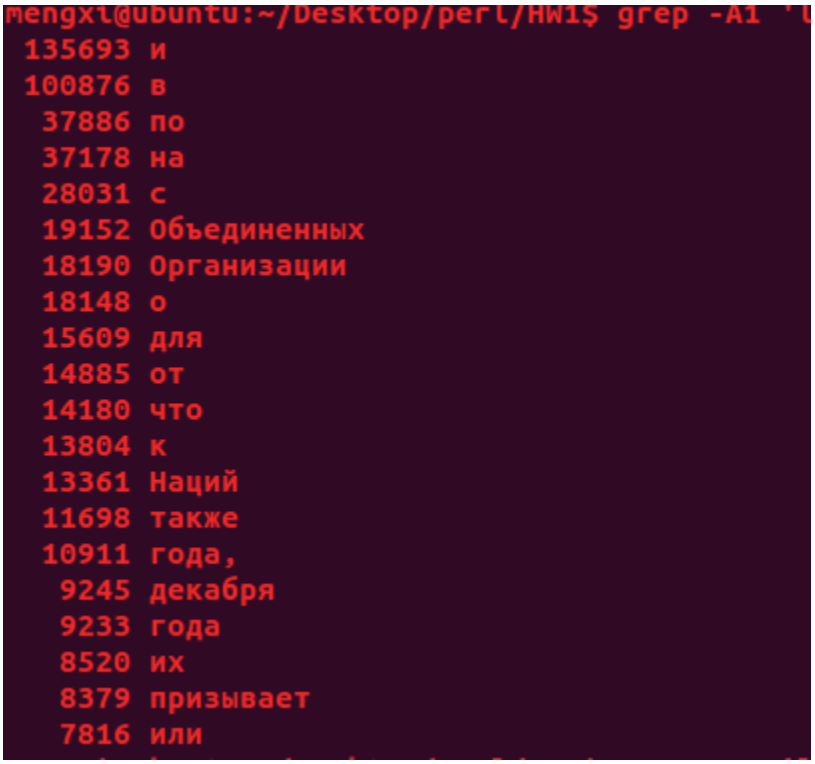
```
mengxi@ubuntu:~/Desktop/perl/HW1$ grep -A1
 267940 the
 175497 of
 136607 and
  99545 to
  66802 in
  35910 on
  32327 for
  22534 that
  21181 its
  20349 with
  20126 a
  20006 United
  19973 as
  17251 by
  16995 Nations
  13933 at
  12827 all
  12061 international
  11986 States
  11173 their
```

b) *grep -A1 'lang="AR"' uncorpora_plain_20090831.tmx | grep -o -P '(?<=<seg>).*(?=</seg>)'| grep -o -P '(?<=\s|\A)[^\s]+(?=\s|\Z)' | sort | uniq -c | sort -nr | head -20*



c) *grep -A1 'lang="ES"' uncorpora_plain_20090831.tmx | grep -o -P '(?<=<seg>).*(?=</seg>)'| grep -o -P '(?<=\s|\A)[^\s]+(?=\s|\Z)' | sort | uniq -c | sort -nr | head -20*

d) *grep -A1 'lang="RU"' uncorpora_plain_20090831.tmx | grep -o -P '(?<=<seg>).*(?=</seg>)'| grep -o -P '(?<=\s|\A)[^\s]+(?=\s|\Z)' | sort | uniq -c | sort -nr | head -20*



B:

1> Arabic

Word	**Meaning**
92683 في	At
45238 من	From
38530 على	On
34817 -	-
34466 إلى	To
24225 أن	That
18986 التي	Which
18947 وإذ	Having, taking
17776 الأمم	Nations
16764 المتحدة	United Nations
15046 عن	about
11324 الدول	States
10979 أو	or
10175 المؤرخ	Of
9476 كانون	Canon
9413 مع	With

9163 جميع	All
8916 بما	including
8800 العام	General
8770 العامة	The public

2> Spanish

Word	**Meaning**
313375 de	From, of
177307 la	La
131113 y	And
93456 en	In
86469 los	Los
82328 el	El
77855 las	Las
77003 a	To
69299 que	Which
52778 del	Of the
37328 para	For
28287 con	With
22750 su	Su
21639 por	For
21407 al	Al
20266 sobre	On
18446 Naciones	Nations
15557 se	Himself, herself, itself
14062 Estados	States
13952 Unidas	United

3> Russian

Word	**Meaning**
135693 и	And
100876 в	In
37886 по	By
37178 на	Per
28031 с	Sec
19152 Объединенных	United
18190 Организации	Organizations
18148 о	O
15609 для	For
14885 от	From
14180 что	That
13804 к	To

13361 Наций	Nations
11698 также	Also
10911 года,	Years,
9245 декабря	December
9233 года	Years
8520 их	Of them
8379 призывает	Encourages
7816 или	for

I think the prepositions and proper noun are similar in various languages. However, because of the methods of translating and language convention, some top words are different in different languages.