Introduction au traitement automatique des langues
TP2 : Tokenisation et découpage en phrase

Partie 1 :

1 - le nombre de tokens (incluant les mots hors vocabulaires) contenus dans la concaténation des corpus du répertoire corpus_topic :

```
aghilas@aghilas:~/Documents/TAL$ cat ./corpus_topic/*/document*.txt | ./tokeniz
e -lex lex_lefff_pos_lemme_freq_90k.txt -line | grep -v "<SAUT-LIGNE/>" | wc -l
310682
```

2 - le taux de mots hors-vocabulaire (tokens n'appartenant pas au lexique) sur toutes les occurrences de tokens de ces corpus :

```
aghilas@aghilas:~/Documents/TAL$ cat ./corpus_topic/*/document*.txt | ./tokeniz
e -lex lex_lefff_pos_lemme_freq_90k.txt -line -oov | grep -v "<SAUT-LIGNE/>" |
wc -l
48281
```

3 - la liste triée par fréquence décroissante de ces formes hors-vocabulaire:

```
aghilas@aghilas:~/Documents/TAL$ cat ./corpus_topic/*/document*.txt | ./tokenize -l
ex lex_lefff_pos_lemme_freq_90k.txt -line -oov | sort | uniq -c | sort -n -r |more
   1136 Le
    950 1
    849 2
    754 Les
    724 3
    665 La
    654 n
    630 0
    616 ne
    605 L
    571 4
    510 5
    457 6
    429 7
    424 Il
    391 Etats
    361 Unis
    312 14
    312 000
    300 t
    285 15
--Plus--
```

4 - les listes classées par fréquence décroissante des formes verbales (étiquette *v*) :

```
aghilas@aghilas:~/Documents/TAL$ grep '       v       ' lex_lefff_pos_lemme_freq_90
k.txt > lex_verbe.txt
aghilas@aghilas:~/Documents/TAL$ cat ./corpus_topic/*/document*.txt | ./tokenize -l
ex lex_verbe.txt -line -no_oov | sort | uniq -c | grep -v "<SAUT-LIGNE/>" | sort -n
 -r | more
```

La même chose pour les formes nominales est les noms propres.

partie 2 :

1 : les listes classées par fréquence décroissante des formes verbales (étiquette *v*):

```
import sys

l1 = sys.stdin.readline()
l2 = sys.stdin.readline()
print("<s>", end=" ")
while l2 :
 print(l1[:-1], end=" ")
 if (l1 == '.\n') :
 if (l2 == "<SAUT-LIGNE/>\n" or (l2[0] >= 'A' and l2[0] <=
'Z')) :
 print("</s>")
 print("<s>", end=" ")
 l1 = l2
 l2 = sys.stdin.readline()
print("</s>")
```

2- Le nombre de phrases obtenu est :

```
aghtlas@aghtlas:~/Documents/TAL$ cat corpus_topic/*/document*.txt |
./tokenize -lex lex_lefff_pos_lemme_freq_90k.txt -line | python3 phrases.py |
wc -l
9691
```

partie 3 :

1 - Un programme pour supprimer les majuscule au début de la phrases :

```
import sys
import os

l = sys.stdin.readline()
while l :
    m = l.split(" ")
    if(m[1] == "<SAUT-LIGNE/>") : pos = 2
    else : pos = 1
    if(m[pos][0] >= 'A' and m[pos][0] <= 'Z') :
        m_min = m[pos].lower()
        cmd = "grep '\t" + m_min + "\t' lex_lefff_pos_lemme_freq_90k.t
        res = os.popen(cmd).read()
        if(res != "0\n") :
            m[pos] = m_min
    for i in range(1, len(m)-2):
        print(m[i], end=" ")
    print()
    l = sys.stdin.readline()
```

2 - Le nombre de mots hors-vocabulaire:

```
aghilas@aghilas:~/Documents/TAL$ cat corpus_topic/*/document*.txt | ./tokenize -lex
 lex_lefff_pos_lemme_freq_90k.txt -line | python3 phrases.py | python3 minuscule.py
 | ./tokenize -lex lex_lefff_pos_lemme_freq_90k.txt -line -oov | grep -v "<SAUT" |
grep -v "LIGNE/>" | wc -l
/bin/sh: 1: Syntax error: Unterminated quoted string
/bin/sh: 1: Syntax error: Unterminated quoted string
/bin/sh: 1: Syntax error: Unterminated quoted string
/bin/sh: 1: Syntax error: Unterminated quoted string
/bin/sh: 1: Syntax error: Unterminated quoted string
/bin/sh: 1: Syntax error: Unterminated quoted string
42974
```