Q10

```
cat uncorpus.eng.txt |grep -o '\b[A-Za-z0-9\.,]\{1,\}[A-Za-z0-9]\b' |
sort|uniq -c|sort -nr
```

In this question, we assumes that words are consisted of A-Z and a-z, numbers are consisted of 0-9, ',', and '.'. But the last letter or digits must be A-Za-z0-9

The Top four sets are:

```
$ cat uncorpus.eng.txt |grep -o '\b[A-Za-z0-9\.,]\{1,\}[A-Za-z0-9]\b'
| sort|uniq -c|sort -nr |head -40
268152 the
176014 of
138293 and
99762 to
66988 in
36024 on
32461 for
24039 that
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
```

5776 Convention

The bottom four sets are:

```
$ cat uncorpus.eng.txt |grep -o '\b[A-Za-z0-9\.,]\{1,\}[A-Za-z0-9]\b'
| sort|uniq -c|sort -nr |tail -40
   1 1,438,826
   1 1,421.6
   1 1,413,400
   1 1,412,370
   1 1,370,050
   1 1,316,374
   1 1,297,600
   1 1,286,710,550
   1 1,277,600
   1 1,267,844,600
   1 1,265,280,800
   1 1,250
   1 1,213,381,487
   1 1,212,554
   1 1,154,600
   1 1,153.9
   1 1,149,200
   1 1,140,446
   1 1,133,672,200
   1 1,119,200
   1 1,117,005
   1 1,100
   1 1,079,000
   1 1,076,225
   1 1,072,500
   1 1,061,400
   1 1,050,000
   1 1,049
   1 1,032,000
   1 1,032
   1 1,025
   1 1,020,000
   1 1,007,000
   1 091
   1 0223
   1 00
   1 0.969
   1 0.4
   1 0.05
   1 0.026
```

The top four sets and the bottom four sets are quite different. The bottom four sets are all numbers. I am not surprised. Because each number will be a unique token.