统计单词出现频率

> 对于给定的英文文本文件，统计各个单词出现的频率，并且显示频率最高的n个单词及其出现次数。
>
> ## 算法步骤
>
> - 将所有大小写将所有大小写英文字母之外的字符转化为一个回车并将多个回车压缩为一个
> - 将所有大写字母转换为小写
> - 排序，结果是按字典顺序排好的单词列表，相同的单词挨在一起
> - 在每个单词之前显示此单词出现的次数
> - 按照出现的次数降序排序
> - 取前n 行输出
> - 输出时加上序号
>
> ## 脚本代码

```
if [ $# -eq 1 ]
then
I_TOP =10
I_FILE=$1
fi

if [ $# -eq 2 ]
then
I_TOP=$1
I_FILE=$2
fi

tr -sc "[A-Z][a-z]" "[\012*]" < $I_FILE | \
tr "[A-Z]" "[a-z]" | \
sort | \
uniq -c | \
sort -k1 -n -r | \
head -$I_TOP | nl
```

代码解析

- 通过tr命令将各种符号用换行符替换
- tr将所有大写字母改为小写
- 排序、统计单词个数
- uniq -c去除重复且在去除之前输出出现的个数
- sort -k1(第一列输出序号) -n出现的次数 -r 降序输出
- head取前10个进行输出

运行结果

> 以rcf2460.txt为测试数据统计词频

```
     1        755 the
     2        346 of
     3        305 header
     4        231 to
     5        195 a
     6        187 ipv
     7        178 in
     8        160 and
     9        157 packet
    10        138 is
```