一、词性频次统计

● 问题描述

给定经过分词、词性标注的中文文本，对中文词性频次进行统计。

● 示例

中文：新华社/NR 乌鲁木齐/NR $date/JJ 电/NN

具体问题：在整篇文档中，统计出每个中文词汇被标注成不同词性的次数
如 新华社 被标注成 NR 出现 10 次，被标注成 NN 出现 1 次…

● 数据

当前文件夹 sample-data\test1 目录下

● 输出示例

```
"官" 的词性有：  3种，分别为：
VV：1次
NN：4次
AD：1次

"在心" 的词性有：    1种，分别为：
VV：1次

"宜" 的词性有：  3种，分别为：
NN：4次
AD：2次
JJ：1次

"实" 的词性有：  6种，分别为：
JJ：3次
VA：1次
AD：3次
VV：2次
NN：4次
DER：    1次
```

二、词汇翻译概率

● 问题描述

给定经过分词的源语言(中文)文本、目标语言文本(英文)、以及词对齐(源语言->目标语)文本，计算词汇翻译概率

● 示例

前提：对a进行map-key处理，便于源语言和目标语言形成对应。

把源语言和目标语言作为key，做词频统计，统计值对应value。
然后，还得对源语言进行词频统计，作为分母用，到时get()取值。
最后计算概率时，就好办了。

源语言： A B C D B A A
目标语言： a b c d
词对齐： 0-0 1-1 2-2 3-3

key

具体问题：在整篇文档中，计算源语言词汇翻译到目标语言词汇的翻译概率：

$$\omega(e_i|f_i) = \frac{count(e_i, f_i)}{\sum_{e_i} count(e_i, f_i)}$$

以（A,a）为例，即：

$$\omega(\mathrm{a}|\mathrm{A}) = \frac{count(A, a)}{count(A)} = \frac{1}{3}$$

$$\omega(NULL|\mathrm{A}) = \frac{count(A, NULL)}{count(A)} = \frac{2}{3}$$

● 数据

当前文件夹 sample-data\test2 目录下

● 输出示例

```
A   ->  a:  1/3
A   ->  NULL:   2/3
B   ->  b:  1/2
B   ->  NULL:   1/2
C   ->  c:  1
D   ->  d:  1
```