> **问题：**
>
> 有一组文件（比如100个.txt文件）,每个文件中的内容为中文、英文、西班牙文、法文四种文字的字或者单词的混排.实际的例子有字幕文件、同声传译文件等要求：使用MapReduce:并行计算框架完成所有文件中每种文字的单词个数的统计，即中文字出现了多少个（总数），英文单词出现了多少个（总数）,西班牙文单词出现了多少个（总数）。

1.Map

读取输入文件,按行读取内容。

- 输入是文件内容,判断每个单词的语言类型(中文、英文、西班牙文、法文)。
- 进行分割,得到每个单词。键是每个单词,值是1
- 输出是<语言类型, 1>。即<中文, 1>,<英文, 1>,<西班牙文, 1>,<法文, 1>。

2.Shuffle

进行数据的分区、排序、聚集

- Partition:
  根据语言类型哈希分区,保证同语言的数据进入同一个reducer。
- Sort:
  在每个分区内部按语言类型排序。
- shuffle:
  将排序结果分发给对应的reducer。

3.Reducer

- 统计每个语言类型的单词数量。
- Emit(语言类型, 总数)。