待处理的文本:

File1.txt

File2.txt

File3.txt

File4.txt

* InputFormat处理主要包含两个过程：
  + 文件 🡪 切片 : getSplits()
  + 切片 🡪 (K,V) : 生成一个RecordReader对象
* InputFormat这个类是一个抽象基类，它的下面有很多子类，而我们用于切片的是FileInputFormat，如果你要自定义一个切片类，那么你就得继承FileInputFormat这个类。如果你不自定义切片类，那么hadoop有一个默认的类TextInputFormat来进行切片工作。
* InputFormat的继承

TextInputFormat 🡪 FileInputFormat 🡪 InputFormat

**下一步 ： InputFormat处理**

* 切片方法：
  + FileInputFormat
    - 切片方法：遍历Input文件夹下的文件🡪获取文件大小🡪计算切片大小(默认blocksize)🡪文件大小小于blocksize则一个文件一片，剩下的部分大于块的1.1倍则两块，否则一块，切片时不考虑数据集整体，而是逐个针对每一个文件单独切片
    - KV方法：
      * key是存储当前行在整个文件中的起始字节偏移量，LongWritable类型；
      * Value是当前行的内容，不包括任何终止符（换行符和回车符），Text类型
  + TextInputFormat
    - 切片方法：FileInputFormat的切片方法
    - KV方法：FileInputFormat的KV方法
  + KeyValueInputFormat
    - 切片方法：FileInputFormat的切片方法
    - KV方法：KeyValueLineRecordReader，每一行分成两部分，第一部分为Key，第二部分为Value。分隔符可以自己指定
  + NLineInputFormat
    - 切片方法：自定义,N行一个切片
    - KV方法：LineRecordReader
  + CombineTextInputFormat
    - 切片方法：针对大量小文件的切片：
    - KV方法：CombineFileRecordReader 和 LineRecordReader几乎一样，只不过涉及到跨文件
* FixedLenthInputFormat：
* 切片方法：FileInputFormat的切片方法
* FixedLenthRecordReader,每次读取的不是一行，而是固定长度
* SeuqenceFileInputFormat
* 切片方法：FileInputFormat的切片方法
* KV方法：SeuqenceFileInputFormat，