si

当生成一个新的segment时（执行flush、commit、merge、addIndexes(facet)），会生成一个描述段文件信息（segmentInfo）的.si索引文件。

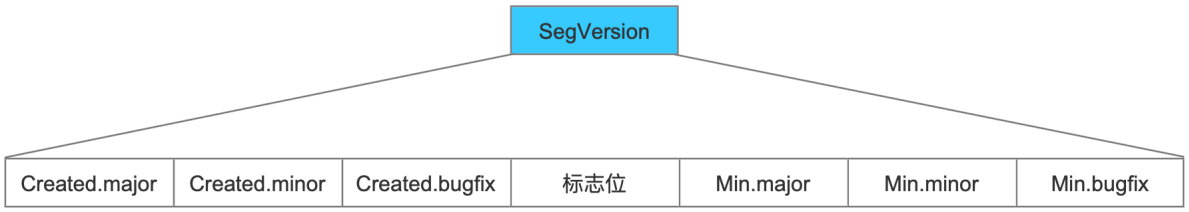
si文件的数据结构

图1：

SegVersion

SegVersion描述了该segment的版本信息。

图2：



Created.major、Created.minor、Created.bugfix

Created描述的是segment创建版本。major、minor、bugfix三者组成了一个版本号，比如本文介绍的就是Lucene7.5.0版本，所以major = 7、minor = 5、bugfix = 0。

标志位

在读取.si文件时的读取该标志位，如果该值为1，表示.si文件中带有Min.major、Min.minor、Min.bugfix的信息需要读取，否则标志位的值为0。

Min.major、Min.minor、Min.bugfix

上文中提到生成一个新的segment可能由flush、commit、merge、addIndexes(facet)触发，那么当merge触发时，意味着多个segment会合并为一个新的segment，那么将某个最小创建版本的segment作为Min.major、Min.minor、Min.bugfix，可以用来判断是否兼容该最小版本的索引文件。

SegSize

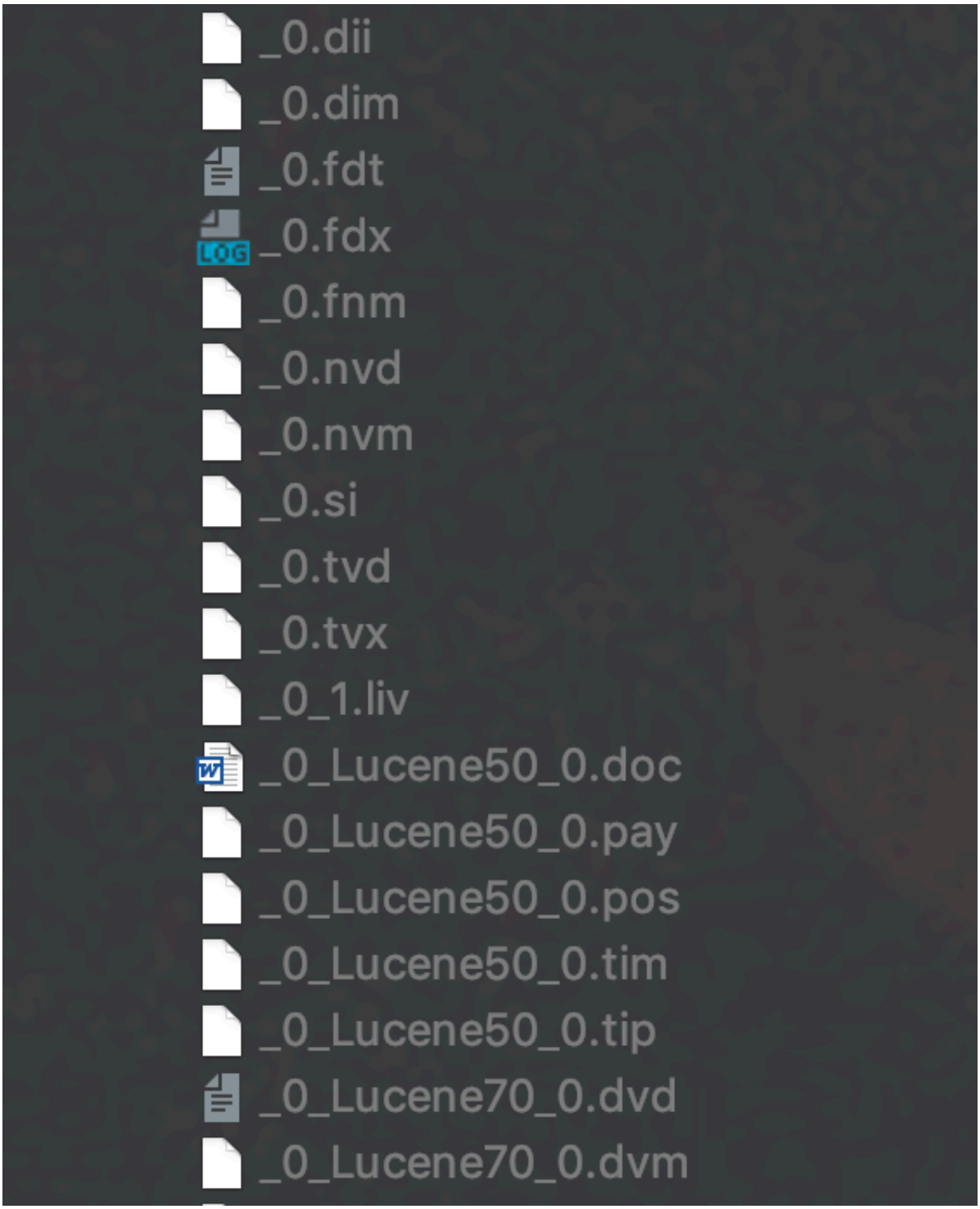
该字段描述了segment中的文档（Document）个数。

IsCompoundFile

该字段描述了segment对应的索引文件是否使用组合文件，在索引文件中生成不同的文件

不使用组合文件会生成.fdx、.fdt、.tvd、tvx、.liv、.dim、.dii、tim、.tip、.doc、.pos、.pay、nvd、.nvm、.dvm、,dvd：

图3：



使用组合文件：

图4：

Diagnostics

该字段描述了以下信息：

- os：运行Lucene所在的操作系统，版本号，架构，比如操作系统为Mac OS X，版本号为10.14.3，架构为x86_64
- java：java的发行商，版本号，JVM的版本号
- version：Lucene的版本号，比如7.5.0
- source：生成当前segment是由什么触发的，flush、commit、merge、addIndexes(facet)
- timestamp：生成当前segment的时间戳

Files

该字段描述了segment对应的索引文件的名字，索引文件即图3或者图4。

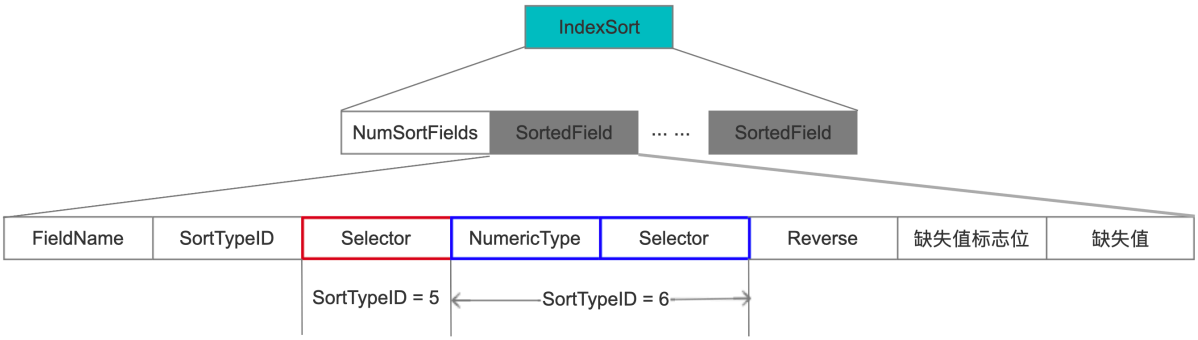
Attributes

该字段描述了记录存储域的索引文件，即.fdx、.fdt使用的索引模式，索引模式有两种:BEST_SPEED 或 BEST_COMPRESSION，Attributes记录其索引模式的名称，两者的区别在 两阶段生成索引文件之第一阶段 已经介绍，不赘述。

IndexSort

该字段用来对segment内的文档进行排序，该值在IndexWriterConfig对象中设置排序规则，可以提供多个Sort对象。该字段会影响文档信息写入索引文件的信息，顺便提一下的是，如果设置了IndexSort后，在 两阶段生成索引文件之第一阶段就只会生成一个临时文件的.fdx、.fdx、.tvd、.tvx文件。

图5：



NumSortFields

该字段描述了排序规则的个数。

FieldName

SortField的域名。

SortTypeID

在前面的文章中介绍了FieldComparator，它同IndexSort一样使用Sort对象来实现排序，而IndexSort中的排序类型（SortField的域值类型）只是FieldComparator中的部分排序类型，每一种排序类型对应一个SortTypeID：

- 0：STRING
- 1：LONG
- 2：INT
- 3：DOUBLE
- 4：FLOAT
- 5：SortedSetSortField
- 6：SortedNumericSortField

Selector

只有SortTypeID = 5时才会有该字段，因为SortedSetSortField允许有多个域值(String类型)，我们必须确定其中一个域值来排序，Selector的值可以有以下几种：

- 0：取最小域值
- 1：取最大域值
- 2：取中间的域值，如果域值个数为偶数个，那么中间的域值就有两个，取较小值，比如有4个域值，"a"，"c"，"d"，"e"，中间域值为"c"，, "d"，那么取"c"
- 3：取中间的域值，如果域值个数为偶数个，那么中间的域值就有两个，取较大值

NumericType、Selector

只有SortTypeID = 6时才会有该字段，因为SortedNumericSortField域值类型NumericType有多个，我们确定域值类型NumericType：

- 0：LONG

- 1：INT
- 2：DOUBLE
- 3：FLOAT

另外因为SortedNumericSortField域值个数可以是多个，所以我们必须确定其中一个域值来排序，Selector的值可以有以下几种：

- 0：取最小域值
- 1：取最大域值

Reverse

该字段为0表示正序，1位倒序。

缺失值标志位

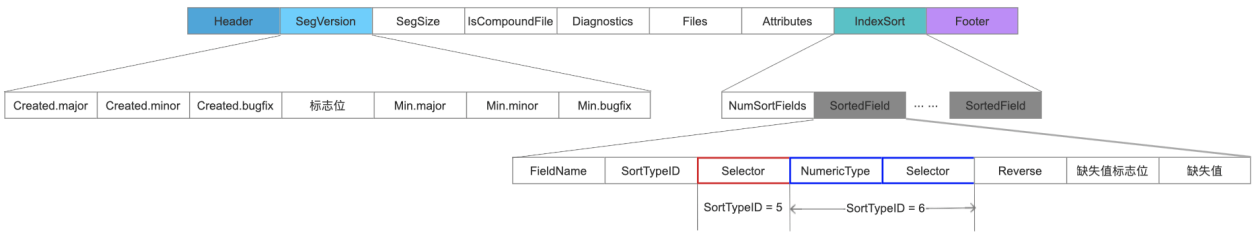
如果有些文档没有排序规则，即需要给该文档添加一个缺失值，那么标志位为1，为0则不添加缺失值。

缺失值

当缺失值标志位为1，那么需要记录缺失值。这里需要特别说明的是，如果域值是String类型，那么它的缺失值只能是固定的 "SortField.STRING_LAST"或者"SortField.STRING_FIRST"，表示没有排序规则的文档要么在序列最后面，要么在序列最前面，其他域值类型需要提供一个明确的缺失值。

si文件总数据结构

图6：



点击下载Markdown文件