**20. 文本分词**

**一、分词简介**

1. 通过数据的清洗，我们可以得到结构比较清晰的文档，这一步需要对文本进行分词。

中文分词指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符，虽然英文也同样存在短语的划分问题，不过在词这一层上，中文比之英文要复杂的多、困难的多。

例如：

研究生命科学

研究生命令本科生

我从马上下来

我马上下来

北京大学生喝进口红酒

在北京大学生活区喝进口红酒

从小学电脑

从小学毕业

美军中将竟公然说

新建地铁中将禁止商业摊点

这块地面积还真不小

地面积了厚厚的雪

让我们以爱心和平等来对待动物

阿美首脑会议将讨论巴以和平等问题

锌合金把手的相关求购信息

别把手伸进别人的口袋里

将信息技术应用于教学实践

信息技术应用于教学中的哪个方面

上级解除了他的职务

方程的解除了零以外还有…

我们一起去故宫

一起恶性交通事故

我不想吃东西

你就不想想

各国有企业相继倒闭

各国有各国的困难

老人家身体不错

老人家中很干净

和服务必归还

技术和服务

他站起身

他起身去北京

问题的确定

这的确定不下来

南京市长江大桥

上面列出来的这几句文字，通过不同的位置分隔，会产生歧义。

2. 常见分词框架对比http://ruby-china.org/topics/28000

[](http://192.168.1.241/doc/exper/129/img/14.png)

①BosonNLP：

http://bosonnlp.com/

http://poem.bosonnlp.com/

②IKAnalyzer，来自Lucene

③结巴分词，小巧功能强大

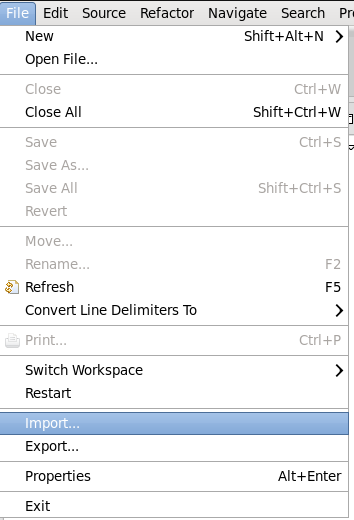
**二、分词项目搭建**

1. 获取分词项目fenci.tar

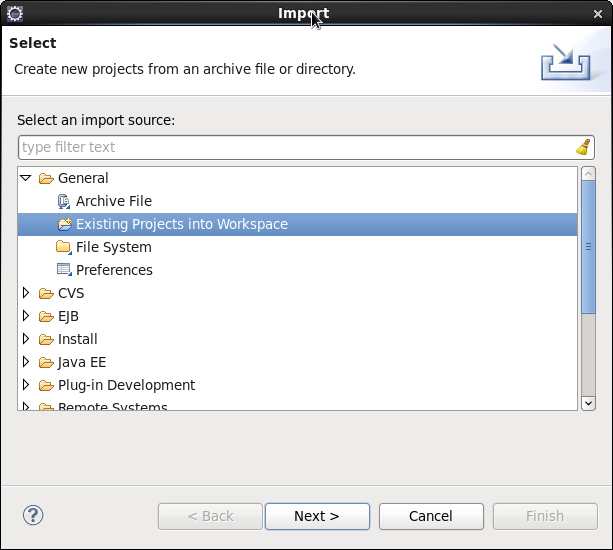
[hfut@master ~]$ cp ~/resources/software/fenci.tar ~/

2. 将解压缩后的项目，导入到eclipse中。

打开eclipse，点击File->Import。

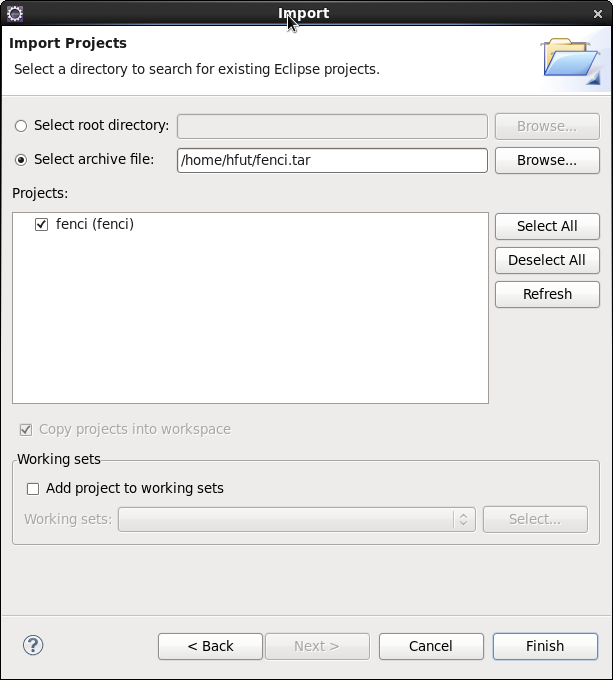


点击导入现有项目到工作空间Existing Projects into workspace



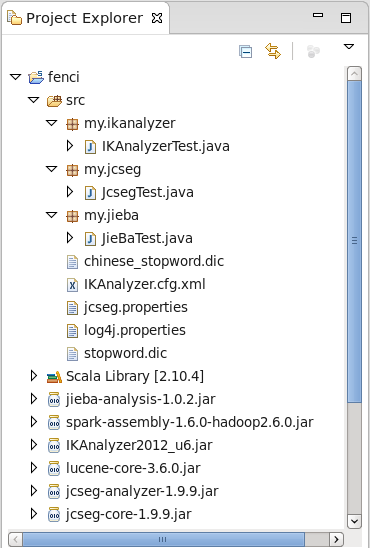
选中Select archive file

选择/home/hfut/fenci.tar文件，点击Finish



3. 执行分词统计操作。

加载完项目后，可以看到项目中，有三种分词的方法，重点来看一下结巴分词和IKAnalyzer。



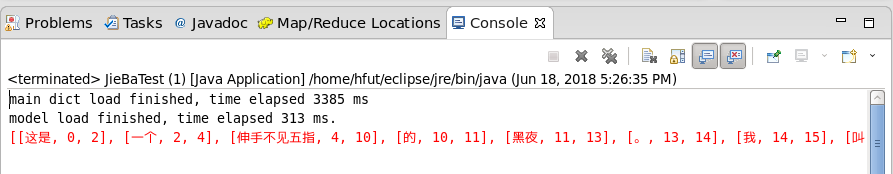
结巴分词：

JiebaSegmenter jieba = new JiebaSegmenter();

String line = "这是一个伸手不见五指的黑夜。我叫孙悟空，我爱北京，我爱Python和C++。";

System.err.println( jieba.process(line, SegMode.SEARCH).toString() );

执行结果：



IKAnalyzer：

String line = "这是一个伸手不见五指的黑夜。我叫孙悟空，我爱北京，我爱Python和C++。";

IKAnalyzer analyzer = new IKAnalyzer(true);

try {

TokenStream tokenStream = analyzer.tokenStream("content", new StringReader(line));

tokenStream.addAttribute(CharTermAttribute.class);

while (tokenStream.incrementToken()) {

CharTermAttribute charTermAttribute = tokenStream.getAttribute(CharTermAttribute.class);

System.out.print(charTermAttribute.toString() + "|");

}

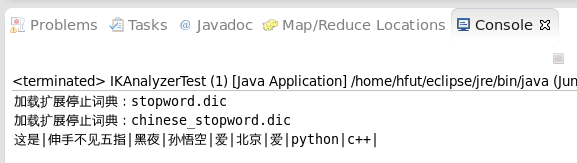
} catch (Exception e) {

System.err.println(e.getMessage());

}

analyzer.close();

执行结果：



4. 执行分词过程中，我们发现很多词语没多大含义。例如英文中的“a the at of”及中文中的“我们、我、一个”。为了提高效率，我们可以把这一类此作为停用词去掉，也就是不记录在分词结果中。

通过修改fenci8项目中，根目录下IKAnalyzer.cfg.xml文件，可以添加停用词库。在这里的配置文件中，已添加了英文的停用词，所以需要添加中文停用词库chinese\_stopword.dic。

<entry key="ext\_stopwords">stopword.dic;chinese\_stopword.dic;</entry>

5. 为了便于在MapReduce、Spark中调用分词，我们需要将分词的方法序列化

