Tika是一个内容抽取的工具集合(a toolkit for text extracting)。它集成了[POI](http://www.oschina.net/p/poi), [Pdfbox](http://www.oschina.net/p/pdfbox) 并且为文本抽取工作提供了一个统一的界面。其次，Tika也提供了便利的扩展API，用来丰富其对第三方文件格式的支持。  
在当前的0.2-SNAPSHOT版本中， Tika提供了对如下文件格式的支持:

* PDF - 通过Pdfbox
* MS-\* - 通过POI
* HTML - 使用[nekohtml](http://www.oschina.net/p/nekohtml)将不规范的html整理成为xhtml
* OpenOffice 格式 - Tika提供
* Archive - zip, tar, gzip, bzip等
* RTF - Tika提供
* Java class - Class解析由[ASM](http://www.oschina.net/p/asm)完成
* Image - 只支持图像的元数据抽取
* XML

Tika的API十分便捷，核心是Parser interface，其中定义了一个parse方法：  
用stream参数传递需要解析的文件流， 文本内容会被传入handler，而元数据会更新至metadata。  
  
可以使用Tika的ParserUtils工具来根据文件的mime-type来得到一个适当的Parser来进行解析工作。或者Tika还提供了一个AutoDetectParser根据不同的二进制文件的特殊格式 (比如说Magic Code)，来寻找适合的Parser。