TikaBatchUsage

Usage

See TikaBatchOverview for a general design overview of tika-batch.

tika-batch was added to Tika 1.8 as its own package, and it was integrated into tika-app with 1.8 as well.

TikaBatch via tika-app-X.Y.jar

You can see the commandline arguments via the regular "-help" commands. There is a separate section at the end for tika-batch options.

In the current dev version. Tika-app decides if it is in batch mode based on one of two signals:

1. There are only two arguments and the first one is an existing directory
2. -inputDir or -i is specified in the commandline

Once the app knows that it is in batch mode, it converts some of the traditional tika-app commandline arguments for use by org.apache.tika.batch.fs.FSBatchProcessCLI.

Some examples:

*Most basic (with output to a directory called "output"):

```
java -jar tika-app.X.Y.jar <inputDirectory> <outputDirectory>
```

*Specify input and output directories:

```
java -jar tika-app.X.Y.jar -i /mydata/src/dir -o /mydata/output/dir
```

*Set the number of file consumer threads:

```
java -jar tika-app.X.Y.jar -numConsumers 10 -i <inputDirectory> -o <outputDirectory>
```

*Output text instead of xml

```
java -jar tika-app.X.Y.jar -t -i <inputDirectory> -o <outputDirectory>
```

*Use the RecursiveParserWrapper and store text for each document:
```
java -jar tika-app.X.Y.jar -J -t -i <inputDirectory> -o <outputDirectory>
```

*Customize the behavior of Tika through the tika-config.xml configuration file:
```
java -jar tika-app.X.Y.jar -c my-custom-tika-config.xml -J -t -i <inputDirectory> -o <outputDirectory>
```

*Specify jvm args to be used by the child process (prepend a "J" to the regular args):

```
java -jar tika-app.X.Y.jar -JXmx2g -JDlog4j.configuration=log4j.xml -i <inputDirectory> -o <outputDirectory>
```

*Commandline to generate output files for tika-eval...only process those files listed in pdfs_random_50000.csv:

```
java -Dlog4j.configuration=file:log4j_driver.xml -cp "bin/*" org.apache.tika.cli.TikaCLI -JXX:-
OmitStackTraceInFastThrow -JXmx5g -JDlog4j.configuration=file:log4j.xml -bc tika-batch-config-basic-test.xml -i
<input_directory> -o <output_directory> -fileList pdfs_random_50000.csv
```

Example logging config files

tika-batch-sh.zip

Some notes

*The watchdog process will restart the child process unless the child process exits with a "do not restart value"=254. If you want to kill all processing, make sure to kill the parent process and then the child process.

*Make sure to add -JXX:-OmitStackTraceInFastThrow to the child process's commandline arguments so that Java doesn't swallow your stack traces.

TikaBatch Server

Module not yet implemented...want to contribute? This would require hardening the server and creating an example client to be used within TikaBatch FS framework.

TikaBatch Hadoop

Module not yet implemented within Tika project...want to contribute? See TikaInHadoop.