Setup: On the Hadoop server, create the following directories:

- hdfs dfs -mkdir -p /input/
- hdfs dfs -mkdir -p /input/articles

- hdfs dfs -mkdir -p /output/
- hdfs dfs -mkdir -p /output/articles

- hdfs dfs -mkdir -p /libs

Unarchive the contents of Wikipedia-EN-20120601_ARTICLES.tar.gz and upload the nlp-tools file using the following commands:

- tar xzvf Wikipedia**-**EN-20120601_ARTICLES.tar.gz
- hdfs dfs -put articles/* /input/articles/ (assuming you're in a local directory called articles)
- hdfs dfs -put opennlp-tools-1.9.3.tar /libs

5a:

For compilation:

- javac -classpath "$(hadoop classpath):./opennlp-tools-1.9.3.jar" -d output DFMapper.java DFReducer.java DFDriver.java
- jar -cvf DFJob.jar -C output .

Execution:

- hadoop jar DFJob.jar DFDriver /input/articles /output/articles /stopwords.txt

5b:

For compilation:

- javac -classpath "$(hadoop classpath):./opennlp-tools-1.9.3.jar" -d output TFIDFMapper.java TFIDFReducer.java TFIDFDriver.java
- jar -cvf DFJob.jar -C output .

Execution:

- hadoop jar TFIDFJob.jar TFIDFDriver /input/articles /output/TFIDF /input/top_100_words.txt /stopwords.txt