

# PythonWordCount

## WordCount Example in Python

This is the [WordCount](#) example completely translated into [Python](#) and translated using [Jython](#) into a Java jar file.

The program reads text files and counts how often words occur. The input is text files and the output is text files, each line of which contains a word and the count of how often it occurred, separated by a tab. To create some input, take your a directory of text files and put it into DFS.

```
bin/hadoop dfs -put my-dir in-dir
```

Each mapper takes a line as input and breaks it into words. It then emits a key/value pair of the word and 1. Each reducer sums the counts for each word and emits a single key/value with the word and sum.

As an optimization, the reducer is also used as a combiner on the map outputs. This reduces the amount of data sent across the network by combining each word into a single record.

To compile the example, build the Hadoop code and the python word count example:

```
ant
cd src/examples/python
./compile
cd ../../..
```

Note that you need to have jythonc and javac on your path for the compilation to work.

To run the example, the command syntax is:

```
bin/hadoop jar src/examples/python/wc.jar in-dir out-dir
```

The results of the word count will be in out-dir/part-.\*.