



Example: Reading a file and getting words list

- Input: **smcng10.txt**

```
smcng10.txt UNREGISTERED
1 Project Gutenberg Etext South American Geology, by Charles Darwin
2 #17 in our series by Charles Darwin
3
4 Copyright laws are changing all over the world, be sure to check
5 the laws for your country before redistributing these files!!!
6
7 Please take a look at the important information in this header.
8 We encourage you to keep this file on your own disk, keeping an
9 electronic path open for the next readers.
10
11 Please do not remove this.
12
13 This should be the first thing seen when anyone opens the book.
14 Do not change or edit it without written permission. The words
15 are carefully chosen to provide users with the information they
16 need about what they can legally do with the texts.
17
18
19 **Welcome To The World of Free Plain Vanilla Electronic Texts**
20
```



Example: Reading a file and getting words list

1. Text file input step, read your file

The top-left screenshot shows the 'Text file input' configuration window with the 'File' tab selected. The 'Selected files' list contains one entry: '1 \${Internal.Entry.Current.Directory}/../files/smcng10...'. The 'Additional output fields' tab is also visible.

The top-right screenshot shows the 'Text file input' configuration window with the 'Fields' tab selected. The 'Separator' is set to '\$(line.separator)' with a red annotation '\$(line.separator)'. The 'Rownum in output?' checkbox is checked, and 'Rownum by file?' is unchecked.

The bottom screenshot shows the workflow diagram. The 'geo file' step is circled in red. The workflow consists of the following steps: 'geo file', 'Split field to rows', 'remove line', 'keep letters and numbers', 'word_length', 'length > 4', and 'PREVIEW'. Below the workflow is a table showing the output of the 'geo file' step:

#	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	line	String										none	N



Example: Reading a file and getting words list

2. Split field to rows step

Split field to rows

Step name: Split field to rows

Field to split: line

Delimiter: |

Delimiter is a Regular Expression:

New field name: word

Additional fields

Include rownum in output? Rownum fieldname:

Reset Rownum at each input row?

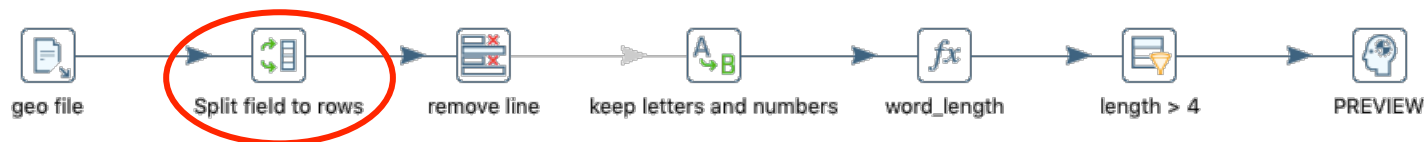
Help OK Cancel

Preview

Examine preview data

Rows of step: Split field to rows (154 rows)

#	line	n	word
1	#17 in our series by Charles Darwin	1	#17
2	#17 in our series by Charles Darwin	1	in
3	#17 in our series by Charles Darwin	1	our
4	#17 in our series by Charles Darwin	1	series
5	#17 in our series by Charles Darwin	1	by
6	#17 in our series by Charles Darwin	1	Charles
7	#17 in our series by Charles Darwin	1	Darwin
8	Copyright laws are changing all over the world, be sure to check	2	Copyright
9	Copyright laws are changing all over the world, be sure to check	2	laws





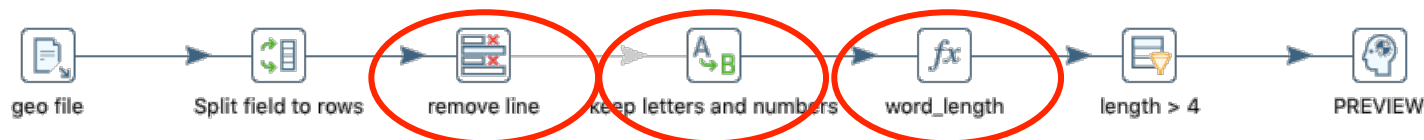
Example: Reading a file and getting words list

3. Add the **Select values** step to remove the **line** field
4. Add **Replace in String** step Remove no relevant words, for example punctuation, by using the regular expression **[^A-Z0-9]** (as alternative, we could use **UDJE** step)
5. Count the number of characters per word

Examine preview data

Rows of step: word_length (154 rows)

#	n	word	word_length
1	1	17	2
2	1	in	2
3	1	our	3
4	1	series	6
5	1	by	2
6	1	Charles	7
7	1	Darwin	6





Example: Reading a file and getting words list

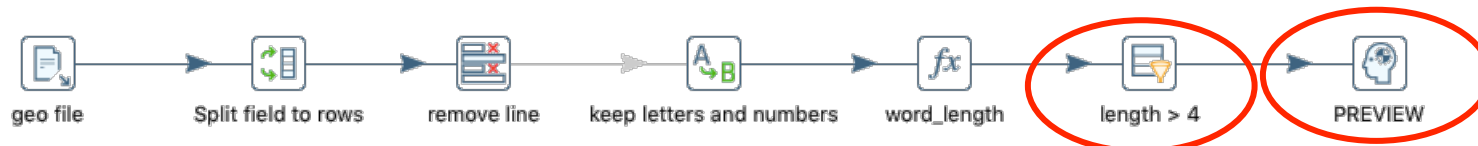
6. **Filter rows** step - a step dedicated to filtering rows based on conditions and comparisons: pass only those rows for which the condition is **true**

word_length >
4 (Integer)

7. Add a **Dummy** step. Creating the hop, select **Main output of step**

Rows of step: PREVIEW (68 rows)

#	n	word	word_length
1	1	series	6
2	1	Charles	7
3	1	Darwin	6
4	2	Copyright	9
5	2	changing	8
6	2	world	5
7	2	check	5





Filter rows step

- We can use more filters:

Condition (example)	Description
word IS NOT NULL	only the rows where the words are neither null nor with empty values pass
line STARTS WITH word	the row passes only if the word field matches the first characters in the line field. Note that in this example the filter involves two fields
word REGEXP (g e).+	pass only the words starting with g or with e
word in list geology; sun	only the rows with words geology and sun pass the filter



Java Filter step

- With the **Java Filter** step you write a Java expression that evaluates to true or false

```
(  
  word CONTAINS un  
  AND  
  len_word>2  
)  
OR  
(word in list world; over)
```



Java expression as

```
(word.matches(".*un.*") && word_length>2) ||  
word.equals("world") || word.equals("over")
```

Examine preview data

Rows of step: combined condition (5 rows)

#	n	word	word_length
1	2	over	4
2	2	world	5
3	3	country	7
4	14	Hundreds	8
5	14	Volunteers	10

