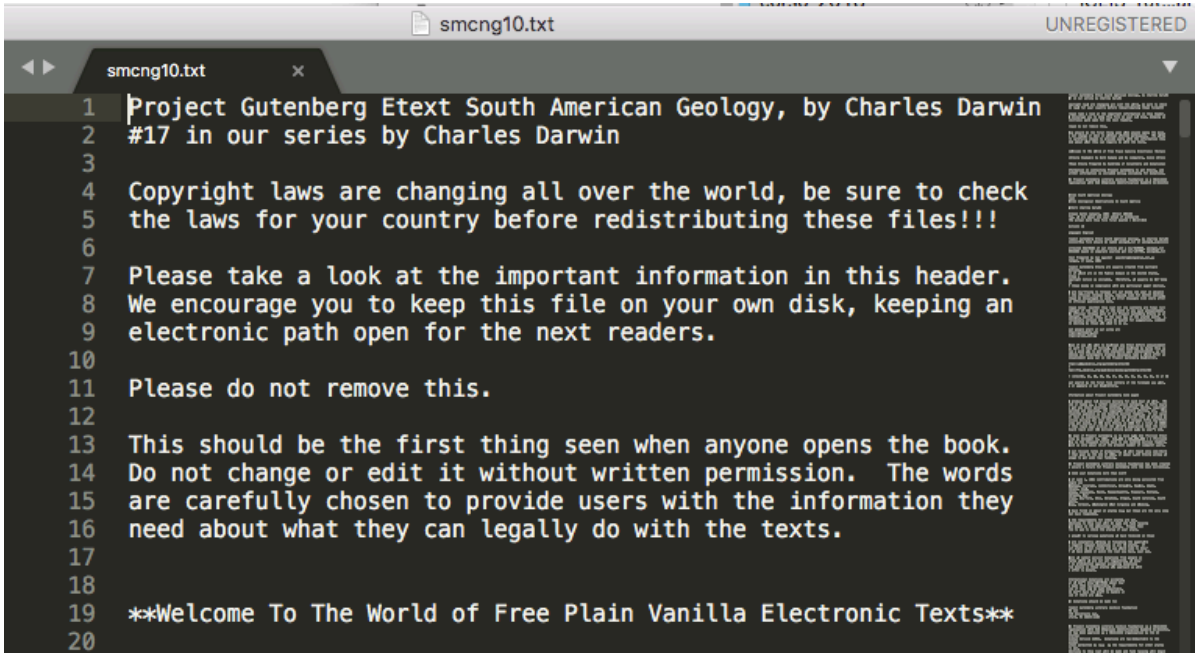
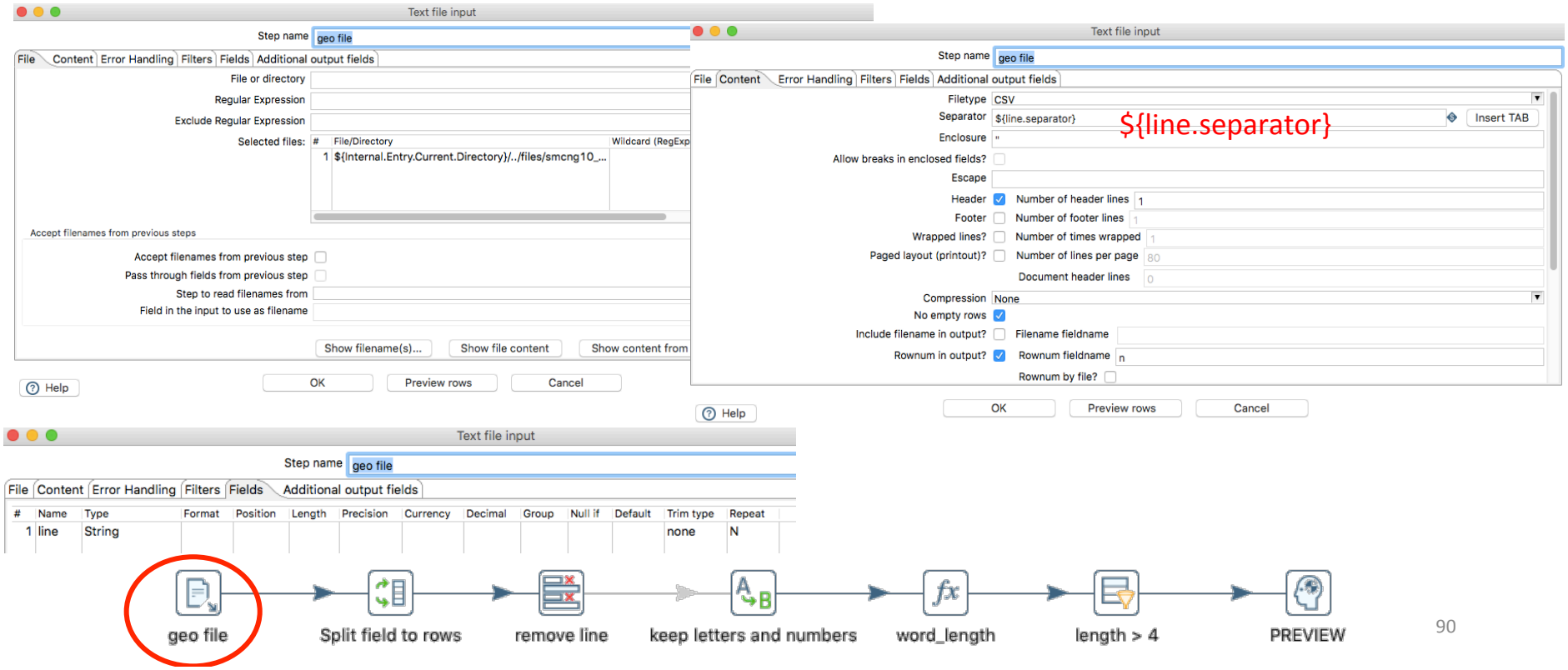
Example: Reading a file and getting words list

- Input: smcng10.txt



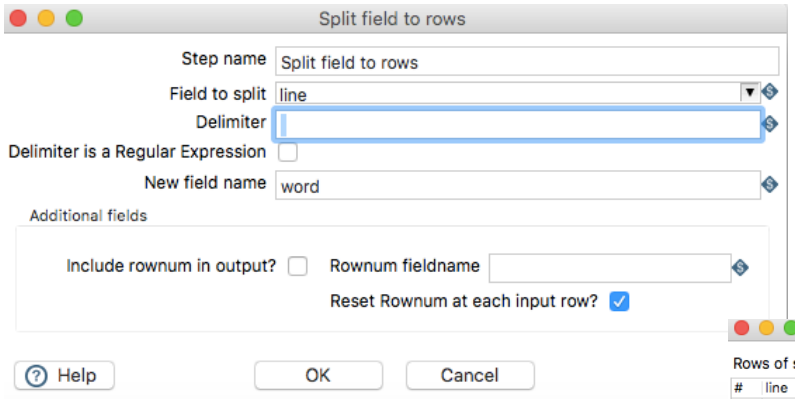
Example: Reading a file and getting words list

1. Text file input step, read your file

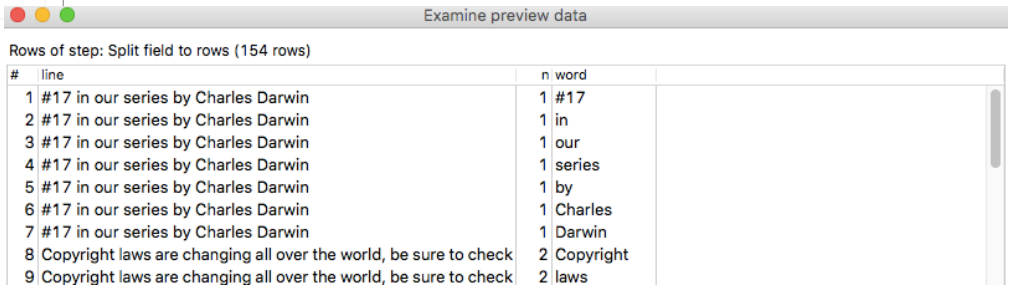


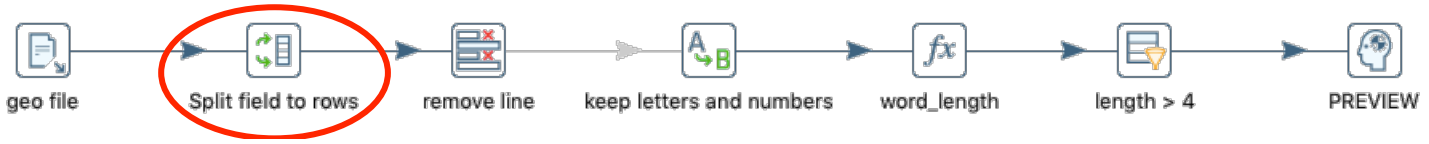
Example: Reading a file and getting words list

2. Split field to rows step



Preview





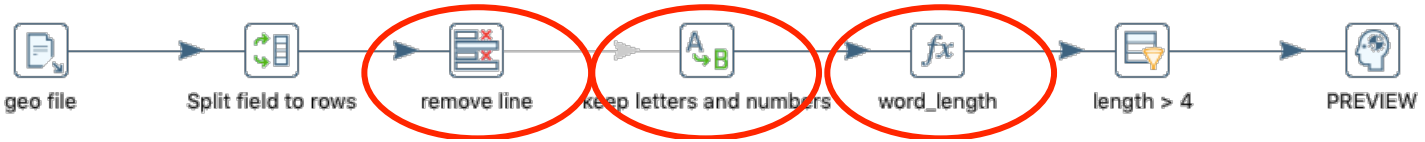
Example: Reading a file and getting words list

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



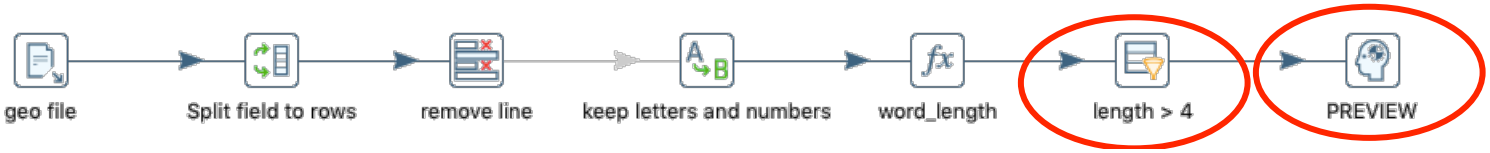
Example: Reading a file and getting words list

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

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