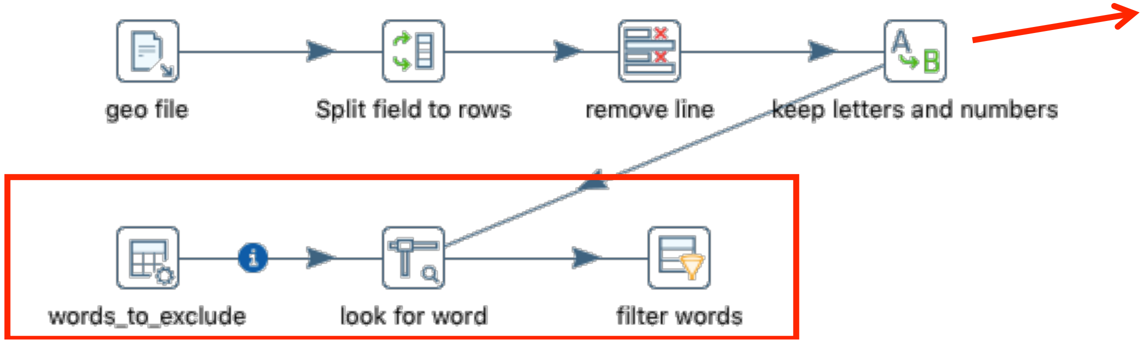
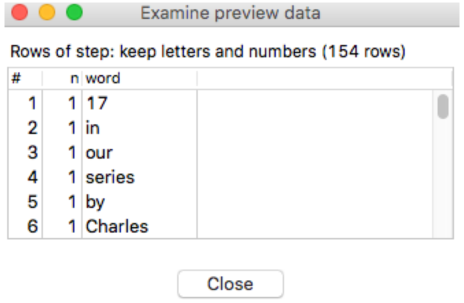
Improving the quality of data

- Example: Reading a file and getting words list
  - We filtered eliminated from the text characters that weren’t part of legal words by using Replace String step
  - Now we want eliminate words in a dictionary, for instance stop words
  - Create a new stream of data that contains words that you want to exclude
    - One word per row, e.g: a, and, as, at, by, from, it, in, of, on, that, the, this, to, which, with, is, are, have, and been.



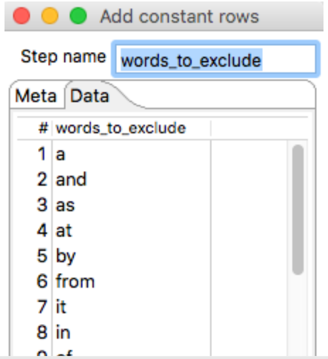


Rows of step: keep letters and numbers (154 rows)

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

Improving the quality of data

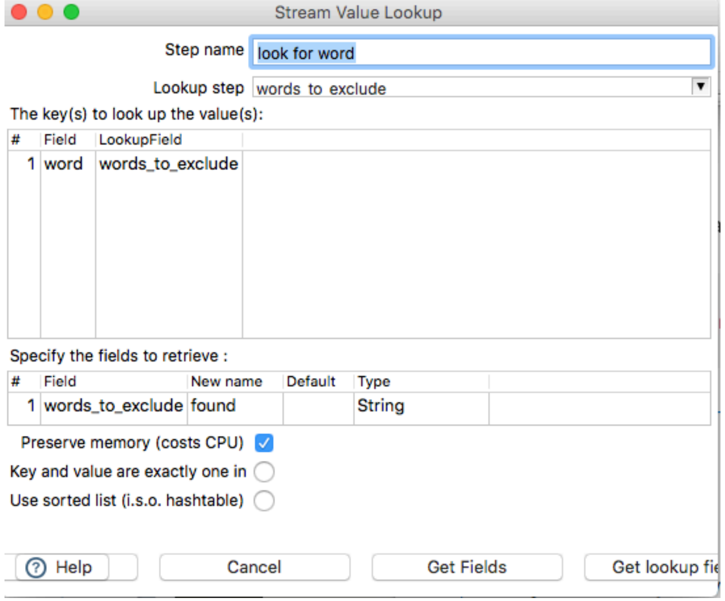
1. Secondary stream: Add Data Grid step or read a list from a plain file
2. Configure Stream Value Lookup step



Add constant rows

Step name: words_to_exclude

#	words_to_exclude
1	a
2	and
3	as
4	at
5	by
6	from
7	it
8	in



Stream Value Lookup

Step name: look for word

Lookup step: words_to_exclude

The key(s) to look up the value(s):

#	Field	LookupField
1	word	words_to_exclude

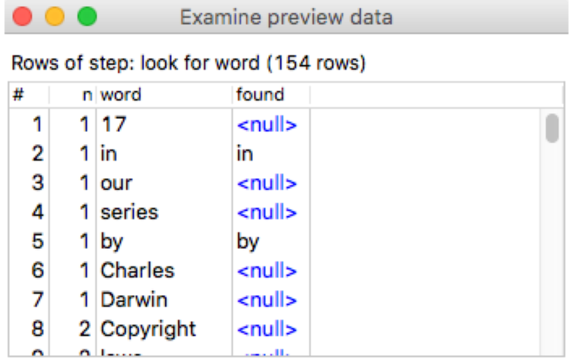
Specify the fields to retrieve :

#	Field	New name	Default	Type
1	words_to_exclude	found		String

Preserve memory (costs CPU) ☑

Key and value are exactly one in ○

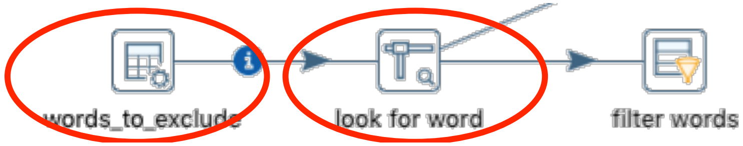
Use sorted list (i.s.o. hashtable) ○



Examine preview data

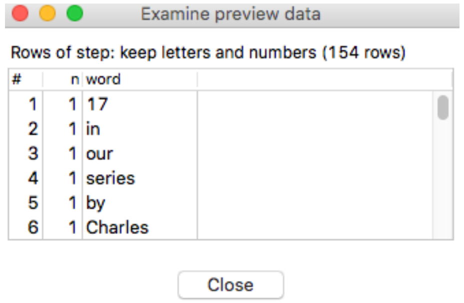
Rows of step: look for word (154 rows)

#	n	word	found
1	1	17	<null>
2	1	in	in
3	1	our	<null>
4	1	series	<null>
5	1	by	by
6	1	Charles	<null>
7	1	Darwin	<null>
8	2	Copyright	<null>



Improving the quality of data

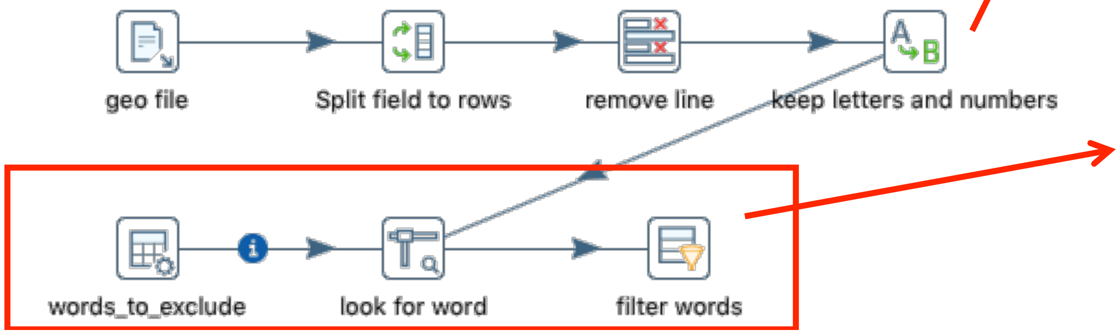
3. Add Filter roes step to discard the common words in the dictionary (found_word IS NULL)

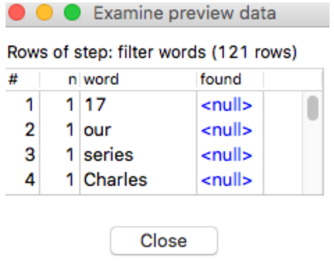


Examine preview data

Rows of step: keep letters and numbers (154 rows)

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





Examine preview data

Rows of step: filter words (121 rows)

#	n	word	found
1	1	17	<null>
2	1	our	<null>
3	1	series	<null>
4	1	Charles	<null>

PDI steps for cleansing data

Step	Description
If field value is null	If a field is null, it changes its value to a constant. It can be applied to all fields of the same data type (e.g. Integer)
Null if...	Sets a field value to null if it is equal to a given constant value
Number range	Creates ranges based on a numeric field (e.g. floating numbers to discrete scale, as 0, 0.25, 0.50, and so on)
Value Mapper	Maps values of a field from one value to another. E.g. yes/no, true/false, or 1/0 values to a unique notation as Y/N
Replace in string	Replaces all occurrences of a string inside a field with a different string (also by using regex)
String operations	e.g. trimming, removing of special characters
Calculator	e.g. remove special characters, convert to upper and lowercase, and retrieve only digits from a string
Stream lookup	Looks up values coming from another stream (e.g. check if in a list or set a default value)
Database lookup	as Stream Lookup, but looks in a database table
Unique rows	Removes double consecutive rows and leaves only unique occurrences