BigML Source Cheat Sheet

Source Configuration Options

Field Types Accepted by BigML

Field Type	Description	Default	Field Icon
Numeric	Used to represent both integer and real numbers.	No limits	123
Categorical (or Nominal)	Used to represent pre-defined values or categories. When BigML processes a field that only takes two values (like 0 or 1), it automatically assigns the type categorical to the field. When BigML detects a field with more than 1,000 categories, it automatically changes the type to text.	1,000 categories per field	ABC
Date-time	Used to represent machine-readable date/time information. When BigML detects a date-time field, by default it expands it into additional fields with their numeric components: year, month, day, day of the week, hour, minute, and second. When disabled, date-time fields will be treated as either categorical or text fields.	No limits	DATE-TIME
Text	Used for text analysis. BigML performs language detection, removes some stop words before processing the text, uses basic stemming, and different tokenization strategies.	1,000 terms per field	text
Items	Used mainly for association discovery. When a field contains an arbitrary number of "items" (categories or labels), BigML assigns the type items to it. Items are separated using a special separator that is configured independently of the separator used to separate the rest of fields of the source.	10,000 items per field	items

Parsing Sources

Option	Description	Default	API Name
Locale	Defines the specific language preferences to process your source and ensure symbols in your data are interpreted in the correct way, e.g., different countries use different symbols for decimal marks.	English (United States)	locale
Multiple Fields source	Detects sources containing multiple fields.	Enabled	separator
Single Field (item-type)	Detects sources containing a single items field when: there are no surrounding quotes, column counts differ from the most frequent count, the proportion of rows is greater than 0.25, there are no missing values as items, there are no items greater in length than 64 characters.	Disabled	separator
Separator	Separates each field within a file. Choose between the following symbols: comma (","), semicolon (";"), tab ("\t"), space (" "), pipe ("|"), or input your own separator.	Comma (",")	separator
Quote	Quotes complete fields, e.g., single quote (') or double quote ("). Adding this symbol is mandatory when the field includes the character used as separator or break lines.	Double quote (")	quote
Missing tokens	Specifies a list of tokens that will be considered as if they were missing values. In addition to the default tokens you can input other tokens.	"", -, ?, NA, NaN, NIL, NULL, N/A, na, null, nil, n/a, #REF!, #VALUE!, #NULL!, #NUM!, #DIV/0, #NAME?, #N/A	missing_tokens
Header	Defines the header information. You can instruct BigML to parse the first line of your CSV file as a header (i.e., the first row is header information) or not (i.e., do not use the first row as header), or let BigML to make that decision for you (i.e., smart header selection).	First row is header information	header
Expand date-time fields	Expands the date-time fields into their numeric components.	Enabled	disable_datetime
Items separator	Selects the specific separator that will be used for items fields: comma (","), semicolon (";"), tab ("\t"), space (" "), pipe ("|"), or other.	Auto detect	separator
Field types	Lets you update the type of each field individually.	Auto detect	fields
Text analysis	Analyzes text fields.	Enabled	term_analysis

Text Analysis

Option	Description	Default	API Name
Language	Chooses the default language of text fields, which will change the resulting stemming, tokenization, and stop words removal. BigML can process text in 21 different languages.	null	language
Tokenization	Chooses whether to split the text into several unique values, treat all the terms in a field as a single value, or apply both modes.	False	token_mode
Stop words removal	Chooses whether or not stop words should be included in the topic model. You can select to remove stop words in the detected language, in all languages or you can keep the stop words	selected_language	stopword_removal
Stop words dilligence	Chooses the aggressiveness of stopword removal, where the levels are light, normal or aggressive, where each level is a superset of words in the previous ones.	normal	stopword_dilligence
Max. n-grams	Chooses the maximum n-gram size to consider for your text analysis. An n-gram is a frequent sequence of *n* terms found in the text. You can choose from a unigram to five-grams	1	ngrams
Stemming	Chooses whether or not lemmatization (stemming) of terms should be applied, according to linguistic rules in the provided language.	True	stem_words
Case sensitive	Chooses whether or not text analysis should be case sensitive.	False	case_sensitive
Filter terms	Chooses to exclude the following groups of terms from your model vocabulary: non-dictionary terms, non-language character terms, numeric digits, HTML keywords, single tokens (to exclude unigrams).	[ ]	filter_terms
Filter specific terms	Specifies the terms that you want to exclude from the model.	[ ]	excluded_terms