BigML Topic Model Cheat Sheet

Topic Model Configuration

Topic Model Configuration Options

Option	Description	Default	API Name
Number of topics	Sets the total number of topics to be found. If it is unset, it will be chosen automatically based on the number documents (i.e., row count). The minimum value is 2 and maximum value is 64.	Unset	number_of_topics
Number of top terms	Sets the total number of the top most influential terms to be shown per topic. The minimum value is 1 and maximum value is 128.	10	top_n_terms
Term limit	Sets the maximum number of total terms in the dataset used by the Topic Model to find the topics.The minimum value is 128 and maximum value is 16384.	4096	term_limit
Min. terms per topic name	Sets the minimum number of top terms to be used for topic names. The minimum value is 0 and the maximum is 10 from the Dashboard.	1	minimum_name_terms
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

Sampling

Option	Description	Default	API Name
Rate	Sets the proportion of the dataset you want to consider between 0% and 100%.	100%	sample_rate
Range	Specifies a subset of instances from which to sample, e.g., from instance 5 to instance 1,000. The **Rate** you set will be computed over the **Range** configured.	(1, max. rows in dataset)	range
Sampling	Allows you to choose between a random sampling or a deterministic sampling. When using deterministic sampling the random-number generator will always use the same seed, producing repeatable results.	Random	seed
Replacement	Allows a single instance to be selected multiple times. Sampling without replacement ensures that each instance cannot be selected more than once.	False	replacement
Out of bag	Selects only the out-of-bag instances for the currently defined sample. If an instance is not selected as part of a sample, it is considered out of bag. It is only selectable when a sample is deterministic and the sample rate is less than 100%.	False	out_of_bag

Batch Topic Distribution Configuration

Output File Options

Option	Description	Default	API Name
Fields separator	Allows you to choose the best separator for your fields.	Comma	separator
Show/hide fields	Allows you to show or hide the rest of the fields in your output file.	True	output_fields
Headers	Allows you to show or hide the names of your columns in the output file.	True	header
New line	Sets the character to use as the line break in the generated csv file: "LF", "CRLF".	LF	newline

Output Dataset

Option	Description	Default	API Name
Output dataset	Defines whether a dataset with the results should be automatically created or not.	True	output_dataset