Topic Modeling with the BigML Dashboard

The BigML Team

Version 2.1

MACHINE LEARNING MADE BEAUTIFULLY SIMPLE

Last updated January 29, 2024

About this Document

This document provides a comprehensive description of how to find relevant topics in text fields using the BigML Dashboard. Learn how to use the BigML Dashboard to configure, visualize, and interpret this unsupervised model and use it to calculate Topic Distributions for single and multiple instances.

This document assumes that you are familiar with:

- Sources with the BigML Dashboard. The BigML Team. June 2016. [6]
- Datasets with the BigML Dashboard. The BigML Team. June 2016. [5]

To learn how to use the BigML Dashboard to build supervised predictive models read:

- Classification and Regression with the BigML Dashboard. The BigML Team. June 2016. [3]
- Time Series with the BigML Dashboard. The BigML Team. July 2017. [7]

To learn how to use the BigML Dashboard to build other unsupervised models read:

- Cluster Analysis with the BigML Dashboard. The BigML Team. June 2016. [4]
- Association Discovery with the BigML Dashboard. The BigML Team. June 2016. [2]
- Anomaly Detection with the BigML Dashboard. The BigML Team. June 2016. [1]

Contents

CHAPTER 1

Introduction

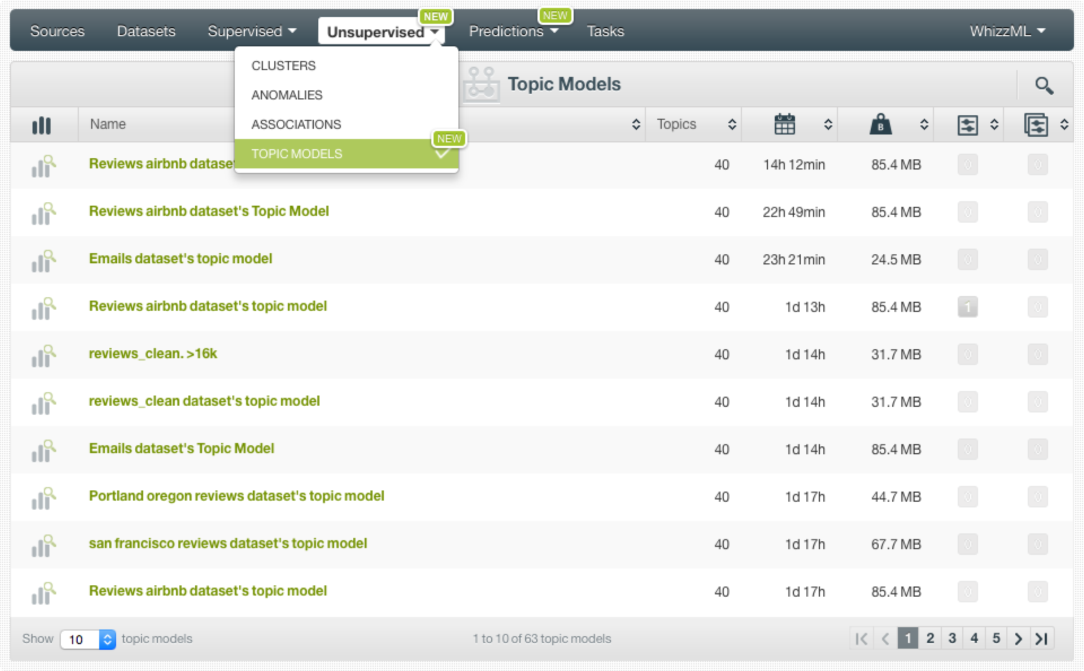
Topic modeling is an unsupervised learning task that helps you find the topics underlying a collection of documents. In other words, it is a form of text mining to identify the hidden thematic structure in a corpus. In order to understand topic models, it is important to take into account three key vocabulary words: **documents**, **topics**, and **terms** where each document is a collection of words, or terms. Topic models assume that documents are generated using one or a combination of topics, and each topic is a group of co-ocurring terms with different probabilities.

The main applications of topic modeling include **browsing**, **organizing** and **understanding** large archives of documents. It can be applied to information retrieval tasks, collaborative filtering, or assessing document similarity among others. Besides these tasks, the output topics can also be very useful as input features for building other models (like classification and regression models, clustering, or anomaly detection).

Topic models in BigML are an optimized implementation of the **Latent Dirichlet Allocation** (LDA) algorithm, one of the best-known probabilistic methods to detect **the relevant topics** within any text, from small text fragments like tweets to long articles, papers or books.

This chapter provides a comprehensive description of BigML topic models, including how they can be created with 1-click (Chapter 3), all the configuration options (Chapter 4), and the visualization provided by BigML (Chapter 5). Once your topic model has been created, you can use it to calculate the topic probabilities for your dataset instances one by one or in batch (Chapter 6). You can even download the topic model to calculate the topic distributions locally (see Section 7.1). You can also create, configure, retrieve, list, update, delete, and use your topic models via the BigML API and bindings (Section 7.2 and Section 7.3).

The fourth tab of the main menu of the BigML Dashboard allows you to select your topic models. The topic model list view (Figure 1.1), shows the **dataset** used to create each topic model, the **Name**, **Topics** (the number of total topics found in the dataset), **Age** (time elapsed since it was created), **Size**, and number of **Topic Distributions** and **Batch Topic Distributions** that have been created using that topic model. The SEARCH menu option in the top right menu of the topic model list view allows you to **search** your topic models by name.



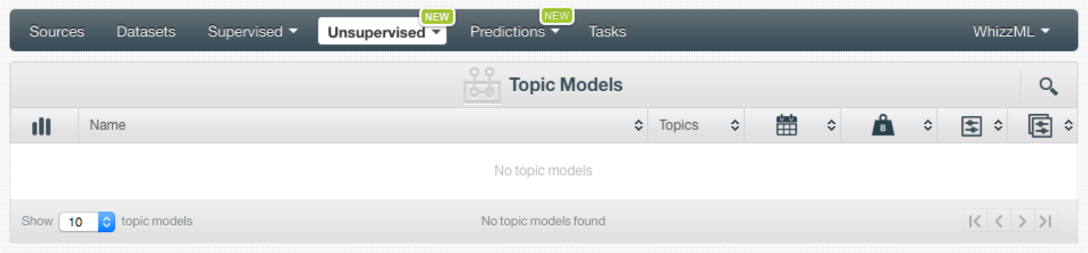
Name	Topics				
Reviews airbnb datase	40	14h 12min	85.4 MB		
Reviews airbnb dataset's Topic Model	40	22h 49min	85.4 MB		
Emails dataset's topic model	40	23h 21min	24.5 MB		
Reviews airbnb dataset's topic model	40	1d 13h	85.4 MB	1	
reviews_clean. >16k	40	1d 14h	31.7 MB		
reviews_clean dataset's topic model	40	1d 14h	31.7 MB		
Emails dataset's Topic Model	40	1d 14h	85.4 MB		
Portland oregon reviews dataset's topic model	40	1d 17h	44.7 MB		
san francisco reviews dataset's topic model	40	1d 17h	67.7 MB		
Reviews airbnb dataset's topic model	40	1d 17h	85.4 MB		

Show 10 topic models

1 to 10 of 63 topic models

Figure 1.1: Topic models list view

When you first create an account at BigML, or every time that you start a new project, your list of topic models will be empty. (See Figure 1.2.)



Name	Topics				

No topic models

Show 10 topic models

No topic models found

Figure 1.2: Topic models empty list view in the BigML Dashboard

Finally, in Figure 1.3 you can see the icon used to represent a topic model in BigML.

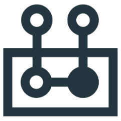


Figure 1.3: Topic model icon

CHAPTER 2

Understanding Topic Models

This chapter describes the BigML topic model resource in detail. For interested users, the chapter will give a deeper understanding of some of the implementation choices and internal mechanisms underlying the BigML topic modeling algorithm.

BigML topic models help you discover the topics underlying a collection of documents. The main goal of topic modeling is finding significant topics to organize, search or understand large amounts of unstructured text data. Topics can also be used as input features for other models, e.g., classification and regression models, cluster analysis or anomaly detection. You can read in this blog post[1] an example of how topic models can be applied to predict the sentiment of movie reviews.

Topic models are based on the assumption that any **document** can be explained as a mixture of **topics**. In BigML, each instance (i.e., each row in your dataset) will be considered a document and the input field, which must be a text field, will be the content of the document. If multiple text fields are given as inputs they will be automatically concatenated so the content for each instance can be considered as a “bag of words”. BigML topic models support any type of text in the following languages; **Arabic, Catalan, Chinese, Czech, Danish, Dutch, English, Farsi/Persian, Finish, French, German, Hungarian, Italian, Japanese, Korean, Polish, Portuguese, Turkish, Romanian, Russian, Spanish, and Swedish**.

BigML topic models are an optimized implementation of Latent Dirichlet Allocation[2] (LDA), one of the most popular probabilistic methods for topic modeling. Since LDA is an unsupervised method, your data does not need to be labeled. Topic models will find the hidden thematic structure underlying the documents.

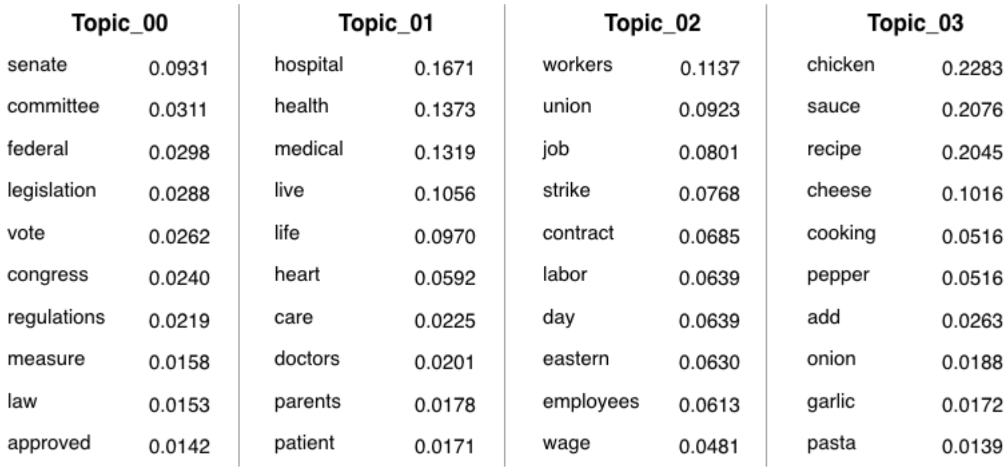
In topic modeling there are three fundamental elements to consider, documents, topics and terms:

- A **document** is a sequence of words of any length. LDA assumes that the terms have been generated using a specific probabilistic distribution over topics. Each document is usually composed of several topics. For example, a document about the impact of new technologies in the schools may be 50% about *technologies*, 30% about *education*, and 20% about *public policy*.
- A **topic** is defined as a group of co-occurring terms which are thematically related. For example, a topic about education may include terms such as “school”, “students”, “elementary”, “children”, etc. The terms from a given topic have different **probabilities** for that topic. All term probabilities for a given topic should sum 100%.
- A **term** is a single lexical token, usually one or more words, but can be any arbitrary string. Each term can be attributable to one or several topics with different probabilities. For example the term “children” may be found in a topic related to education but also in a topic related to clothing.

You can see an example of en English language topic model in Figure 2.1, which shows the topics found in a dataset containing a collection of newspaper articles. Each topic is composed of a group of terms with different probabilities. Note that topics are not labeled, i.e., the topic names do not give any idea

[1] https://blog.bigml.com/2016/11/18/predicting-movie-review-sentiment-with-topic-models/
[2] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

of their content, their meaning needs to be inferred according to the terms they are composed of. By looking at each group of terms in the image below we can interpret the first topic as public policy-related, the second as healthcare-related and so on. A useful way to improve the topic name interpretation is to use the BigML's default where the top N terms are selected as the topic names (see **??**).



Topic_00		Topic_01		Topic_02		Topic_03	
senate	0.0931	hospital	0.1671	workers	0.1137	chicken	0.2283
committee	0.0311	health	0.1373	union	0.0923	sauce	0.2076
federal	0.0298	medical	0.1319	job	0.0801	recipe	0.2045
legislation	0.0288	live	0.1056	strike	0.0768	cheese	0.1016
vote	0.0262	life	0.0970	contract	0.0685	cooking	0.0516
congress	0.0240	heart	0.0592	labor	0.0639	pepper	0.0516
regulations	0.0219	care	0.0225	day	0.0639	add	0.0263
measure	0.0158	doctors	0.0201	eastern	0.0630	onion	0.0188
law	0.0153	parents	0.0178	employees	0.0613	garlic	0.0172
approved	0.0142	patient	0.0171	wage	0.0481	pasta	0.0139

Figure 2.1: Example of topics extracted from a collection of newspaper articles

If you want to learn more, professor David Blei, one of the inventors of LDA together with Andrew Y. Ng and Michael I. Jordan, gives a very insightful tutorial on it here[3].

2.1 General Text Analysis

Any dataset containing text data needs some pre-processing so it can be used to properly train a Machine Learning model. In BigML you can configure text parameters like the **tokenization** strategy, the **case sensitivity**, **stemming** or **stop words** using the Text Analysis options at the **source configuration level** explained in the **Sources with the BigML Dashboard** [6]. These text configuration options define the vocabulary for all BigML models, including topic models. However, topic models are the only models that allow you to change this configuration at the topic model creation time (see Chapter 4). You can build several topic models using different text configurations without going back to the source and having to create different datasets.

Among the text configuration options for topic models you can find parameters like the **language**, the **tokenization**, whether to include or exclude **stop words**, the maximum size for the **n-grams** to be considered in your model vocabulary, the **stemming**, and the **case sensitivity**. You can also define the terms that you want to completely **exclude** from your model. All these options are explained in Section 4.5.

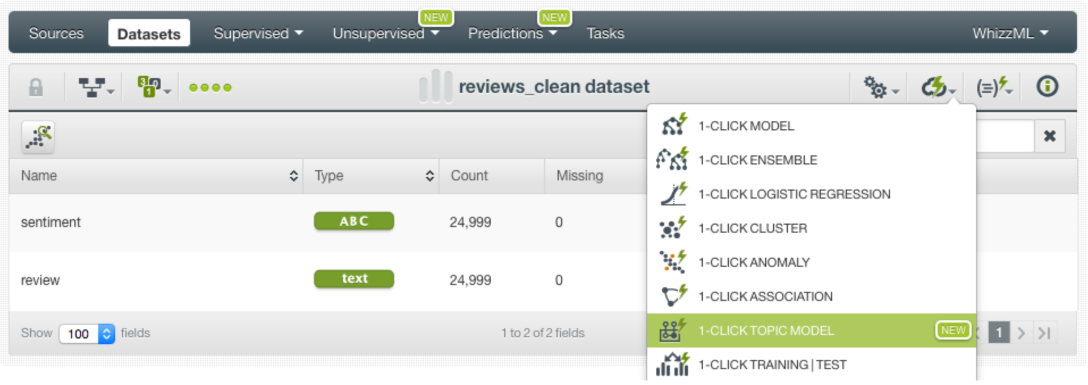
[3] http://videolectures.net/mlss09uk_blei_tm/

CHAPTER 3

Creating Topic Models with 1-Click

To create a topic model in BigML you have two options: you can use the **1-click option** which uses the default values for all available configuration options, or you can tune the parameters in advanced by using the **configuration option** explained in Chapter 4.

You can find the 1-CLICK TOPIC MODEL option in the **1-click action menu** from the dataset view. (See Figure 3.1.)



reviews_clean dataset

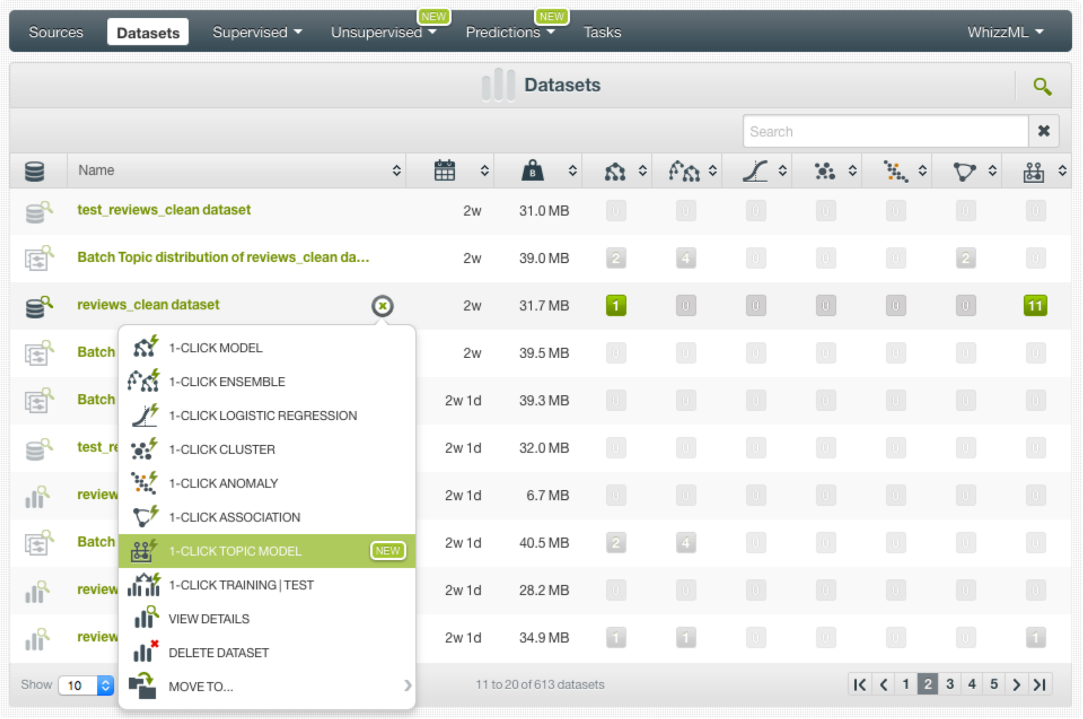
Name	Type	Count	Missing
sentiment	ABC	24,999	0
review	text	24,999	0

Show 100 fields

1 to 2 of 2 fields

Figure 3.1: Create topic model from 1-click action menu

Alternatively, you can use the 1-CLICK TOPIC MODEL option in the **pop up menu** from the dataset list view. (See Figure 3.2.)



Name									
test_reviews_clean dataset	2w	31.0 MB	0	0	0	0	0	0	0
Batch Topic distribution of reviews_clean da...	2w	39.0 MB	2	4	0	0	0	2	0
reviews_clean dataset	2w	31.7 MB	1	0	0	0	0	0	11
Batch	2w	39.5 MB	0	0	0	0	0	0	0
Batch	2w 1d	39.3 MB	0	0	0	0	0	0	0
test_re	2w 1d	32.0 MB	0	0	0	0	0	0	0
review	2w 1d	6.7 MB	0	0	0	0	0	0	0
Batch	2w 1d	40.5 MB	2	4	0	0	0	0	0
review	2w 1d	28.2 MB	0	0	0	0	0	0	0
review	2w 1d	34.9 MB	1	1	0	0	0	0	1

Figure 3.2: Create topic model from popu up menu

Either option builds a topic model using the default values for the available configuration options explained in the following section. (See Chapter 4.)

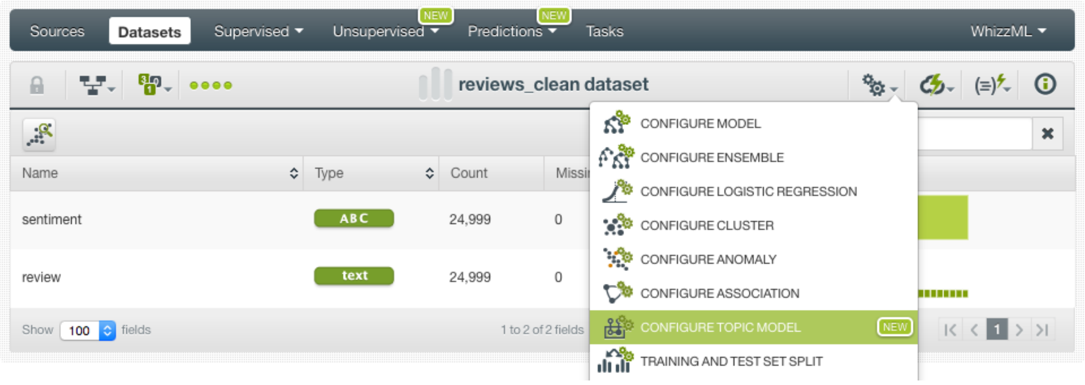
Note: please remember that topic models only support text fields, therefore although your dataset may contain a number of different field types, the topic model will only take into account the text fields and will ignore the rest. If your dataset does not contain any text field, the 1-CLICK TOPIC MODEL option will be disabled.

CHAPTER 4

Topic Model Configuration Options

You can configure a number of parameters that affect the way BigML creates topic models. You can find an explanation of each parameter.

To display the configuration panel to see all options, click the CONFIGURE TOPIC MODEL menu option in the **configuration menu** from the dataset view. (See Figure 4.1.)

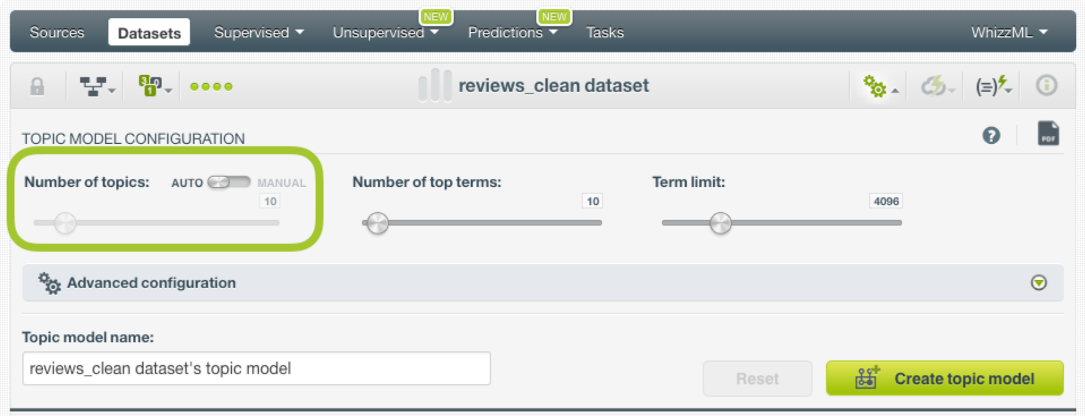


Name	Type	Count	Missi
sentiment	ABC	24,999	0
review	text	24,999	0

Figure 4.1: Configure topic models

4.1 Max. Number of Topics

You can specify the total **number of topics** to be found (MANUAL option) or you can let BigML automatically discover them according to the number of instances in your dataset (AUTO option). By default the AUTO option is active. The maximum number of topics to be discovered is 64. (See Figure 4.2.)



Number of topics: AUTO MANUAL 10

Number of top terms: 10

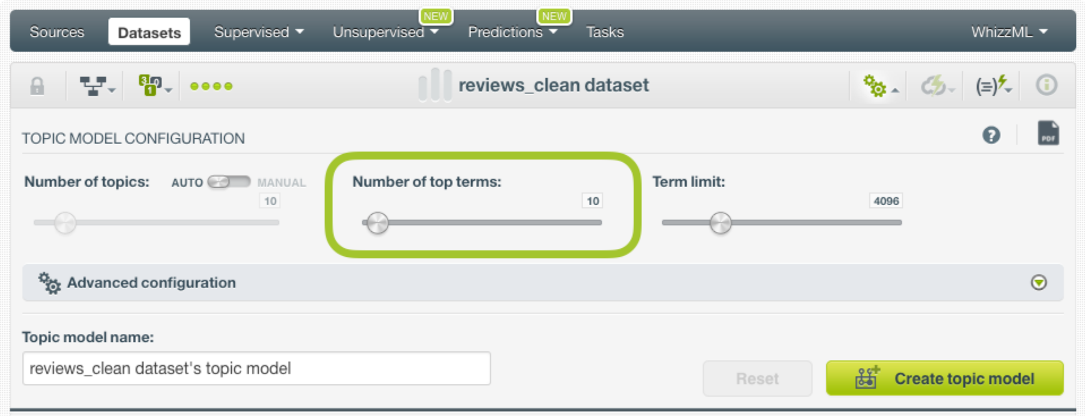
Term limit: 4096

Topic model name: reviews_clean dataset's topic model

Figure 4.2: Configure topic models

4.2 Number of Top Terms

Each topic is composed by a number of different terms with different probabilities. All the term probabilities for a single topic should sum 100%. You can select a fix **number of top terms** per topic (ordered by probability) to be displayed in your topic model view. You can select up to 128 terms. By default BigML shows 10 terms per topic.



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

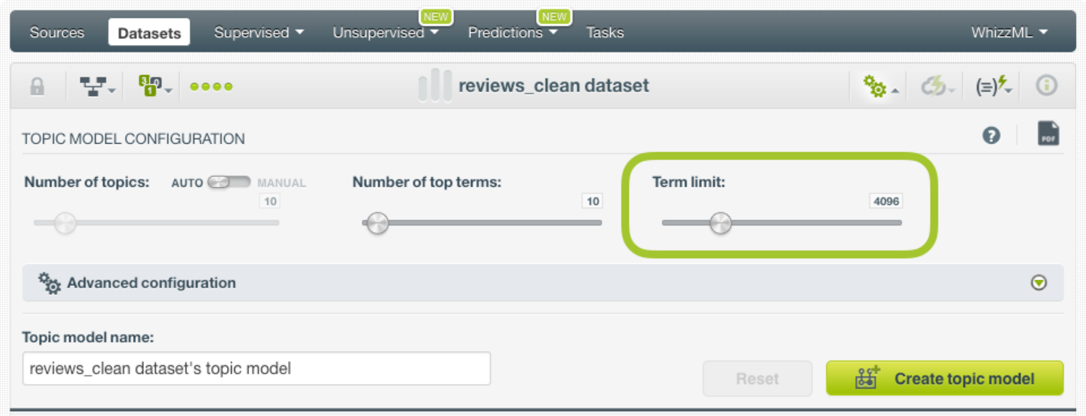
Topic model name: reviews_clean dataset's topic model

Figure 4.3: Select the number of top terms to be displayed in the topic model view

4.3 Number of Terms

Depending on the dataset, the number of total unique terms can be pretty high and most of them may not be relevant to define the boundaries between the topics. Therefore the model builds the vocabulary by setting a maximum number of terms to be considered, ignoring the less relevant ones. The term relevance is measured by the term frequency: those terms that are too frequent or too infrequent can be discarded since they are not expected to be good discriminators between topics. You can set the **term limit** to build the model vocabulary up to 16,384 terms. By default BigML takes the most relevant 4,096 terms.

The ideal number of total terms will depend on the number of topics, the number of instances in your dataset, and the length and homogeneity of your texts. A large collection of long and very diverse documents will need a bigger model vocabulary.



Number of topics: AUTO MANUAL 10

Number of top terms: 10

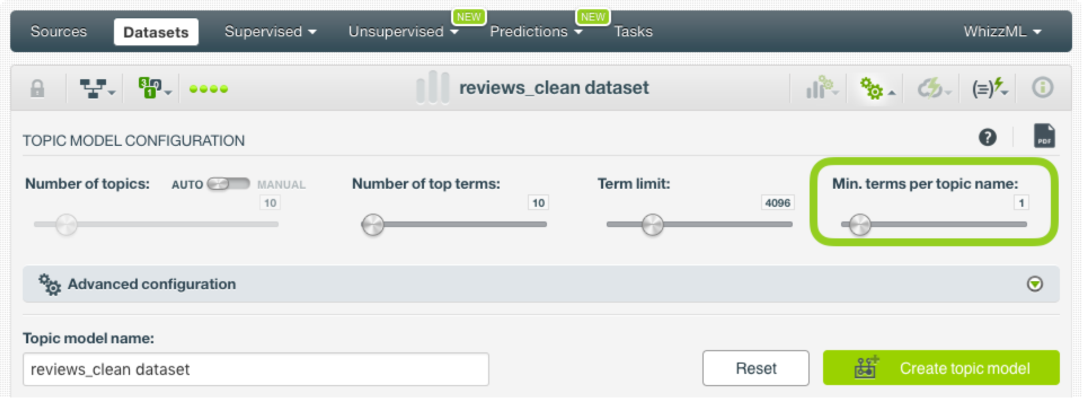
Term limit: 4096

Topic model name: reviews_clean dataset's topic model

Figure 4.4: Select the number of terms to build the model vocabulary

4.4 Minimum Terms Per Topic Name

By default, BigML selects the first term per topic to set that topic name. You can select up to 10 terms to name your topics. If two or more topics have the same top N terms selected, then the next most influencial term will be added to these topic names. If you select 0 terms, BigML will set generic enumerated names per topic ("Topic 00", "Topic 01", "Topic 02", etc.).



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Min. terms per topic name: 1

Topic model name: reviews_clean dataset

Figure 4.5: Select the number of terms to be included in the topic names

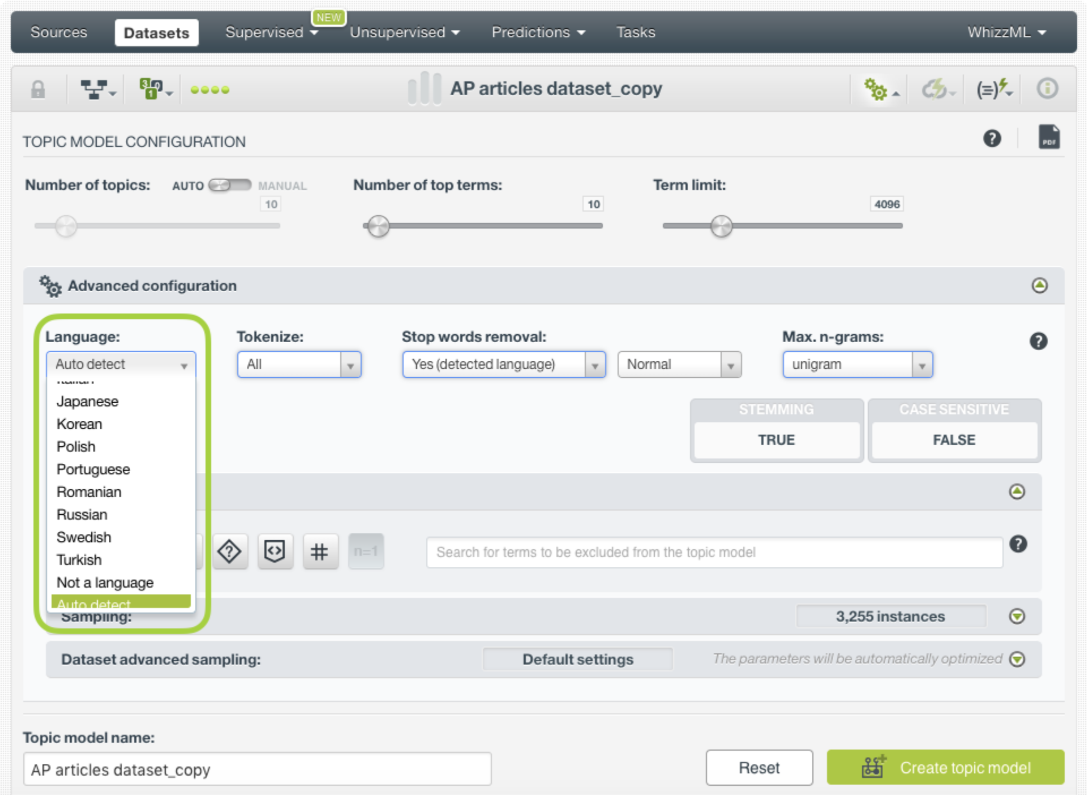
4.5 Text Analysis

As explained in Chapter 2, BigML provides several configuration parameters to process your text fields and build the model vocabulary. See the subsections below for an explanation of each parameter.

Note: topic models take into account the text analysis configuration at the source level. For example, if you configured the source to accept stop words, by default, the topic model will also consider them.

4.5.1 Language

BigML attempts to do basic language detection of each text field. For your topic you can choose any of the following languages: **Arabic, Catalan, Chinese, Czech, Danish, Dutch, English, Farsi/Persian, Finish, French, German, Hungarian, Italian, Japanese, Korean, Polish, Portuguese, Turkish, Romanian, Russian, Spanish, and Swedish**. The selected language will be applied for all the text dataset fields.



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

STEMMING TRUE

CASE SENSITIVE FALSE

3,255 instances

Dataset advanced sampling: Default settings *The parameters will be automatically optimized*

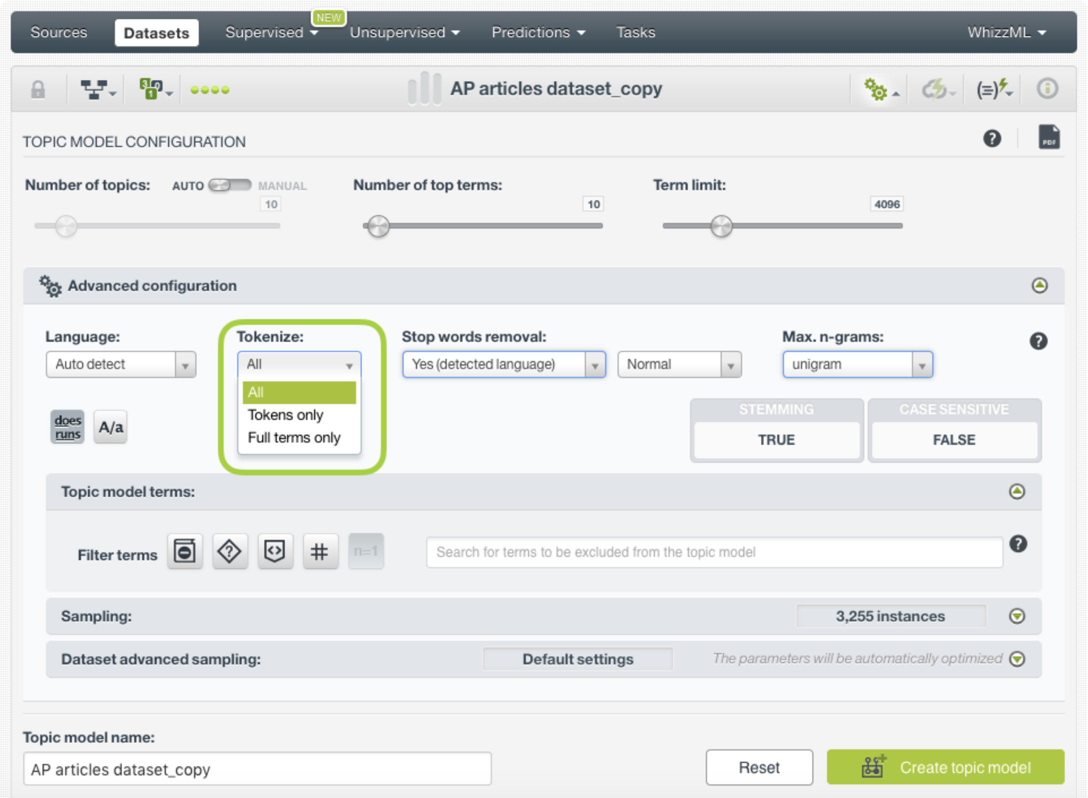
Topic model name: AP articles dataset_copy

Figure 4.6: Language configuration options

4.5.2 Tokenize

Tokenization strategy allows splitting the text into several unique values. You can choose one of the following methods (default is “**All**”):

- **Tokens only**: individual words are used as terms. For example, “ML for all” becomes [“ML”, “for”, “all”].
- **Full terms only**: the entire field is treated as a single term as long as it is shorter than 256 characters. In this case “ML for all” stays [“ML for all”]
- **All**: both full terms and tokenized terms are used. In this case [“ML for all”] becomes [“ML”, “for”, “all”, “ML for all”].



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

All
Tokens only
Full terms only

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

STEMMING TRUE

CASE SENSITIVE FALSE

Sampling: 3,255 instances

Dataset advanced sampling: Default settings The parameters will be automatically optimized

Topic model name: AP articles dataset_copy

Figure 4.7: Tokenize configuration options

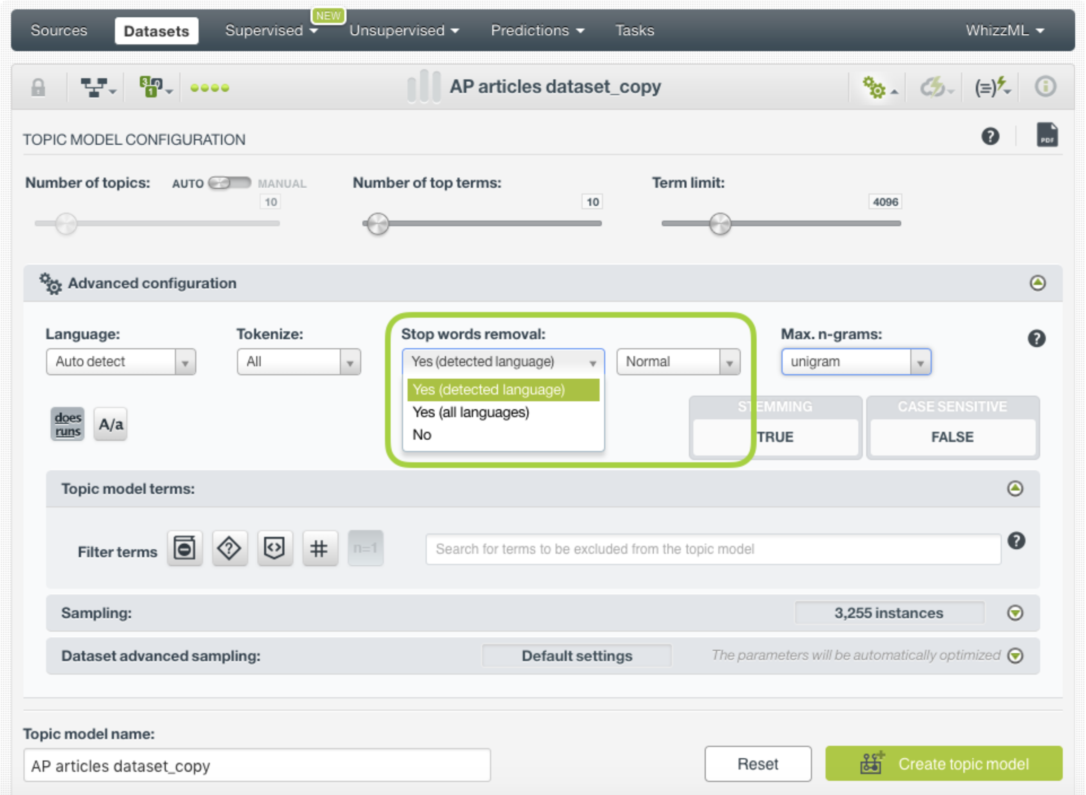
4.5.3 Stop Words Removal

The **Stop words removal** selector allows you to remove the use of usually uninformative stop words[1] as part of the model vocabulary. Some examples of stop words are: **a**, **the**, **is**, **at**, **on**, **which**, etc. Obviously, these change according to the language chosen to process each text field. This is the reason why BigML offers three options:

- **Yes (detected language)**: this option removes the stop words only for the detected language. If you have several languages mixed within the same field or different languages for different fields, the stop words of the non-detected languages will appear in your models. This is the option selected by default.
- **Yes (all languages)**: this option removes the stop words for all languages. Although you have several languages mixed within the same field, you will not find any stop words in your models. The downside is that some stop words for some languages may be valid words for other languages.
- **No**: this option will avoid the stop words removal. Therefore, the stop words will be included in your model vocabulary.

Next to the **Stop words removal** selector you will find another selector that allows you to choose the aggressiveness of stopword removal where each level is a superset of words in the previous ones: **Light**, **Normal**, and **Aggressive**. By default, BigML performs **Normal** stop words removal.

[1] `https://en.wikipedia.org/wiki/Stop_words`



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

STEMMING TRUE

CASE SENSITIVE FALSE

Sampling: 3,255 instances

Dataset advanced sampling: Default settings

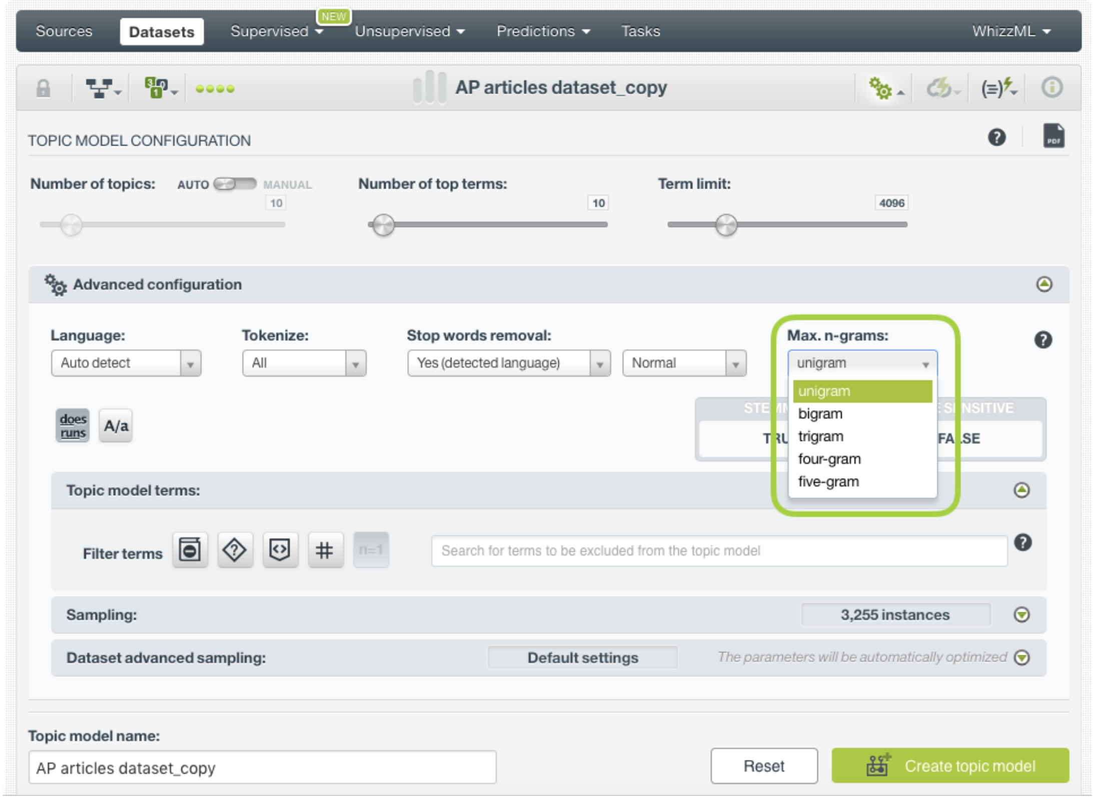
Topic model name: AP articles dataset_copy

Figure 4.8: Stop words configuration options

4.5.4 Max. N-Grams

The **Max. n-grams** selector allows you to choose the maximum n-gram[2] size to consider for your text analysis. An n-gram is a frequent sequence of *n* terms found in the text. For example, “market” is a unigram (n-gram of size one), “prime minister” is a bigram (n-gram of size two), “Happy New Year” is a trigram (n-gram of size three), and so on. If you choose to keep stop words, they will be considered for the n-grams. You can select from unigrams up to five-grams.

[2] https://en.wikipedia.org/wiki/N-gram



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

Sampling: 3,255 instances

Dataset advanced sampling: Default settings

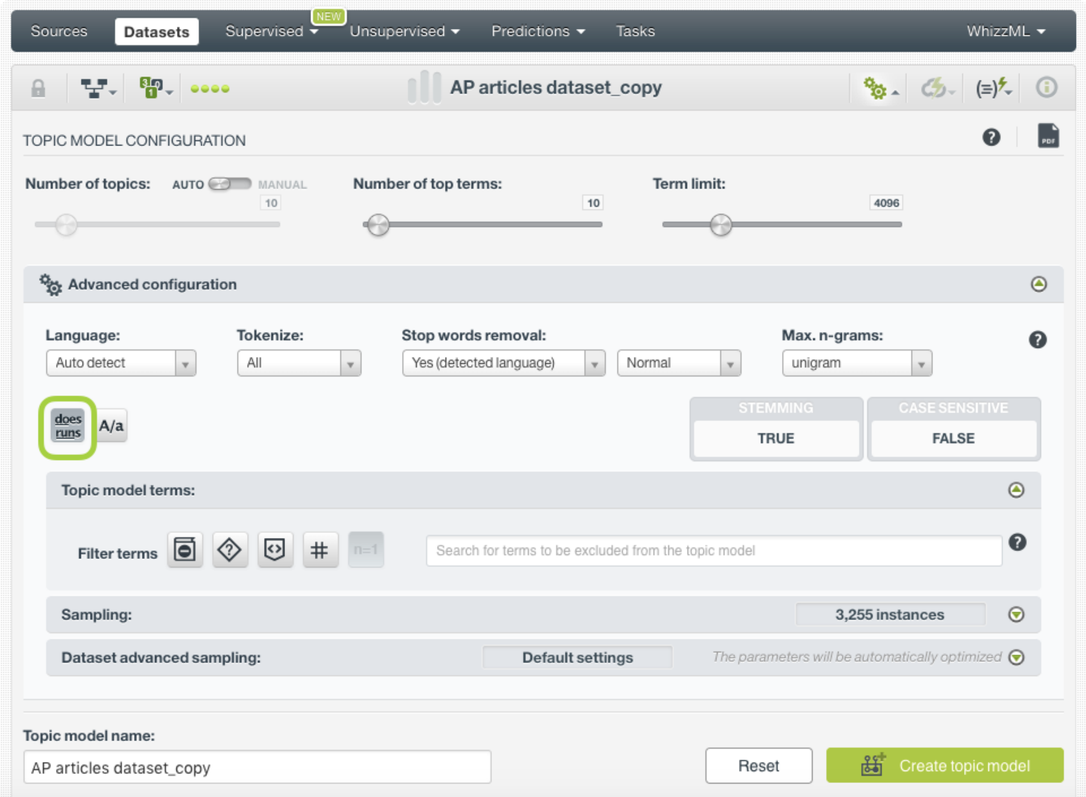
Topic model name: AP articles dataset_copy

Figure 4.9: n-grams configuration options

4.5.5 Stemming

BigML can differentiate all possible words or apply stemming[3], so words with the same root are considered one single value. For example, if stemming is enabled, the words great, greatly and greatness would be considered the same value instead of three different values. This option is enabled by default.

[3] https://en.wikipedia.org/wiki/Stemming



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

STEMMING TRUE

CASE SENSITIVE FALSE

Sampling: 3,255 instances

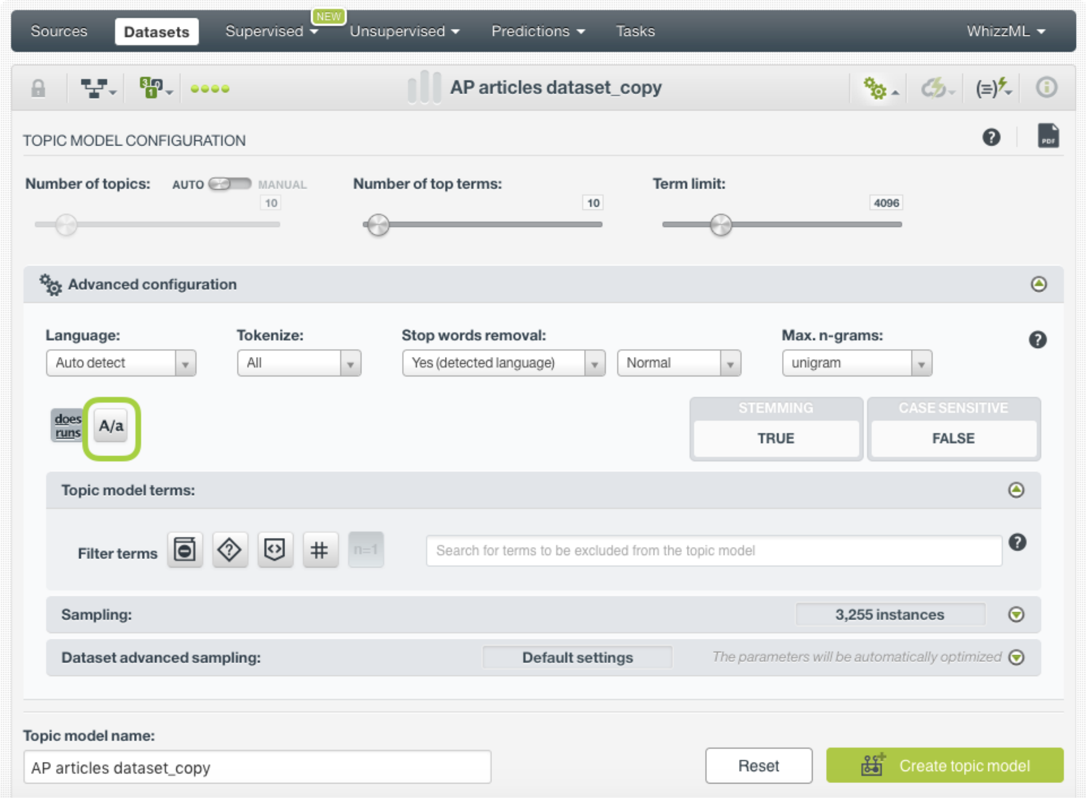
Dataset advanced sampling: Default settings

Topic model name: AP articles dataset_copy

Figure 4.10: Stemming configuration

4.5.6 Case Sensitivity

Specify whether you want BigML to differentiate words if they contain upper or lower cases. If you click the case sensitivity option, terms with lower and upper cases will be differentiated, e.g., “House” and “house” will be considered two different terms. This option is inactive by default.



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

does runs A/a

STEMMING TRUE

CASE SENSITIVE FALSE

Topic model terms:

Filter terms n=1 Search for terms to be excluded from the topic model

Sampling: 3,255 instances

Dataset advanced sampling: Default settings *The parameters will be automatically optimized*

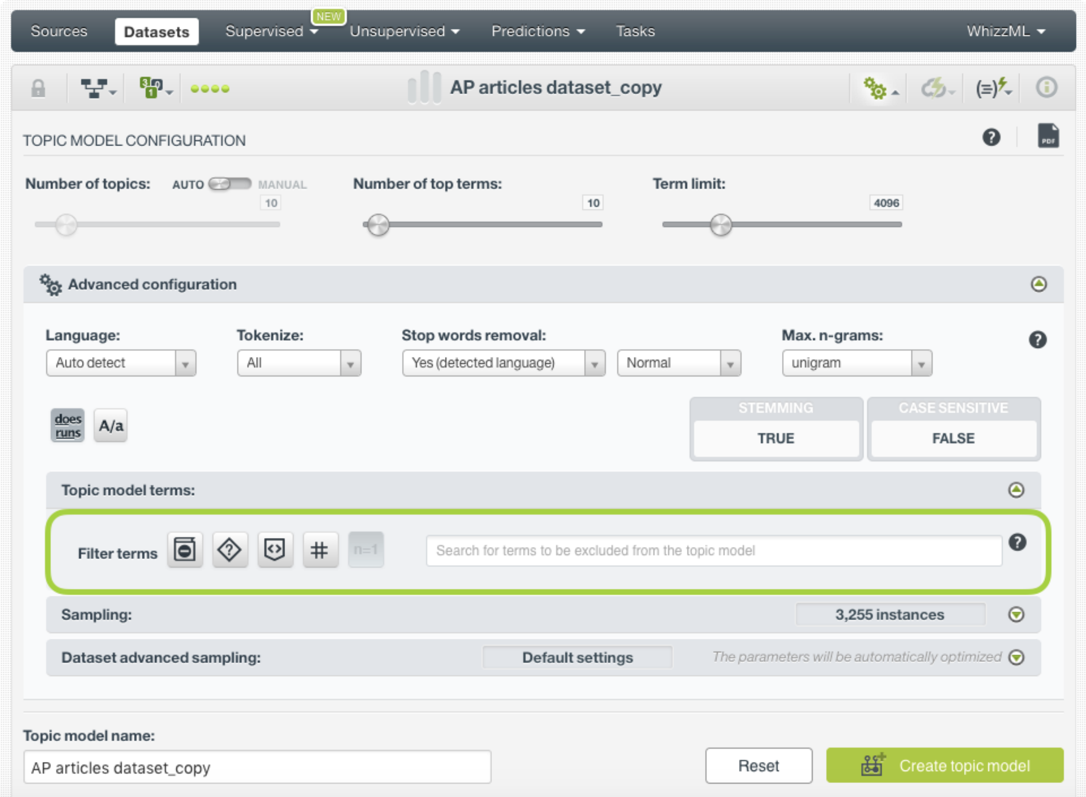
Topic model name: AP articles dataset_copy

Figure 4.11: Case sensitivity configuration

4.5.7 Filter Terms

You can select to exclude certain terms from your model vocabulary. BigML provides the following otpions:

- **Non-dictionary words**: this option excludes terms that are unusual in the provided language. For this filter, BigML uses its own custom dictionaries that are composed of different sources such as online word lists, parses of Wikipedia, movie scripts, etc. These source may change depending on the language. The words in our dictionaries might contain terms like slang, abbreviations, proper names, etc. depending on whether or not these words are common enough to be found in our internet sources.
- **Non-language characters**: this option excludes terms containing uncommon characters for words in the provided language. For example, if the language is Russian, all terms containing non-Cyrillic characters will be filtered out. Numeric digits will be considered non-language characters regardless of language.
- **HTML keywords**: this option excludes JavaScript/HTML keywords commonly seen in HTML documents.
- **Numeric digits**: this option excludes any term that contains a numeric digit in [0-9].
- **Single tokens**: this option excludes terms that contain only a single token, i.e., unigrams. Only bigrams, trigrams, four-grams, five-grams and/or full terms will be considered. At least one of these options needs to be selected, otherwise the single token filter will be disabled.
- **Specific terms**: this is a free text option where you can write any term or group of terms to be excluded from your model. This option is very useful to remove terms which may appear in topics but are not relevant to define their themes.



Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

STEMMING TRUE

CASE SENSITIVE FALSE

Sampling: 3,255 instances

Dataset advanced sampling: Default settings *The parameters will be automatically optimized*

Topic model name: AP articles dataset_copy

Figure 4.12: Filter terms

4.6 Sampling Options

Sometimes you do not need all the data contained in your testing dataset to generate your topic models. If you have a very large dataset, sampling may be a good way of getting faster results. (See Figure 4.13.) You can configure the sampling options explained in the following sections.

4.6.1 Rate

The rate is the proportion of instances to include in your sample. Set any value between 0% and 100%. It defaults to 100%.

4.6.2 Range

Specifies a subset of instances from which to sample, e.g., choose from instance 1 until 200. The **Rate** you set will be computed over the **Range** configured.

4.6.3 Sampling

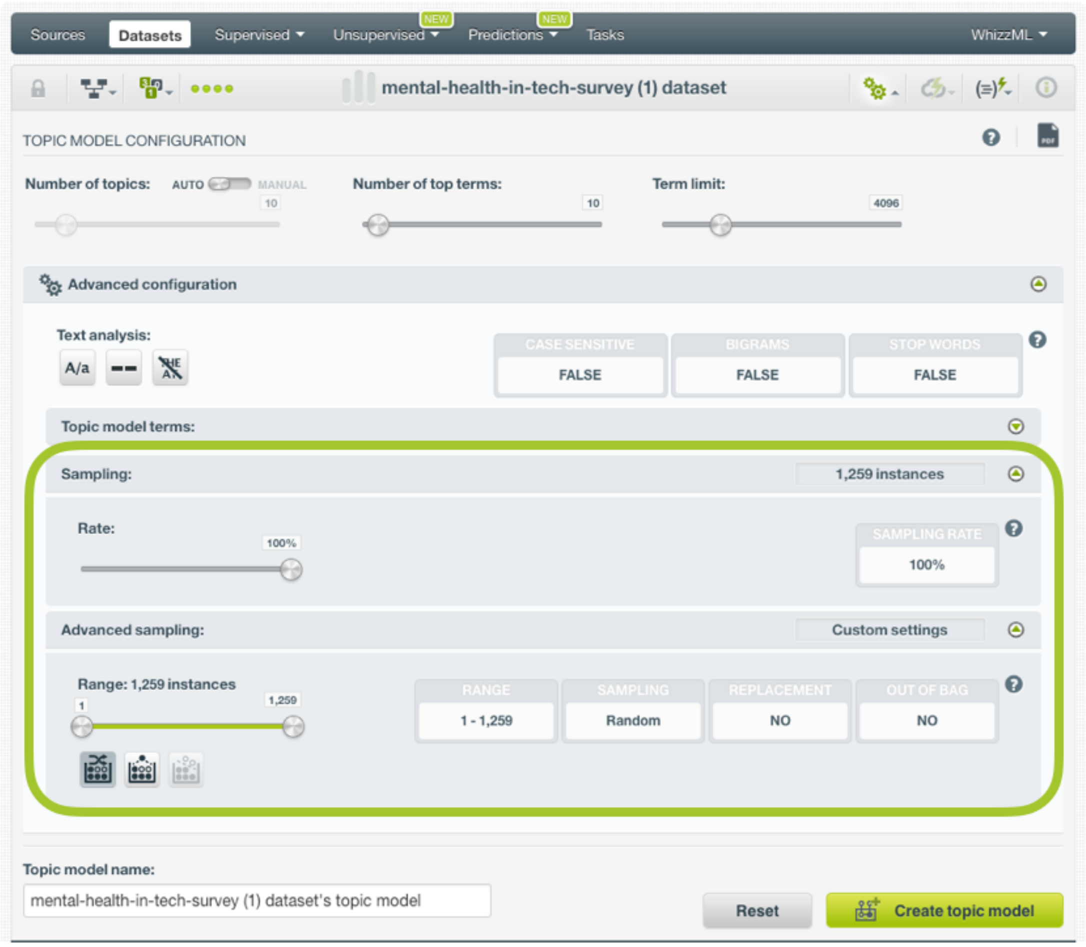
By default, BigML selects your instances for the sample by using a random number generator, which means two samples from the same dataset will likely be different even when using the same rates and row ranges. If you choose deterministic sampling, the random-number generator will always use the same seed, thus producing repeatable results. This lets you work with identical samples from the same dataset.

4.6.4 Replacement

Sampling with replacement allows a single instance to be selected multiple times. Sampling without replacement ensures that each instance cannot be selected more than once. By default, BigML generates samples without replacement.

4.6.5 Out of Bag

This argument will create a sample containing only out-of-bag instances for the currently defined rate. If an instance is not selected as part of a sample, it is considered out of bag. Thus, the final total percentage of instances for your sample will be 100% minus the rate configured for your sample (when replacement is false). This option can only be selected when a sample rate is less than 100%.



mental-health-in-tech-survey (1) dataset

TOPIC MODEL CONFIGURATION

Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Text analysis: CASE SENSITIVE FALSE; BIGRAMS FALSE; STOP WORDS FALSE

Topic model terms:

Sampling: 1,259 instances

Rate: 100%; SAMPLING RATE 100%

Advanced sampling: Custom settings

Range: 1,259 instances 1 1,259

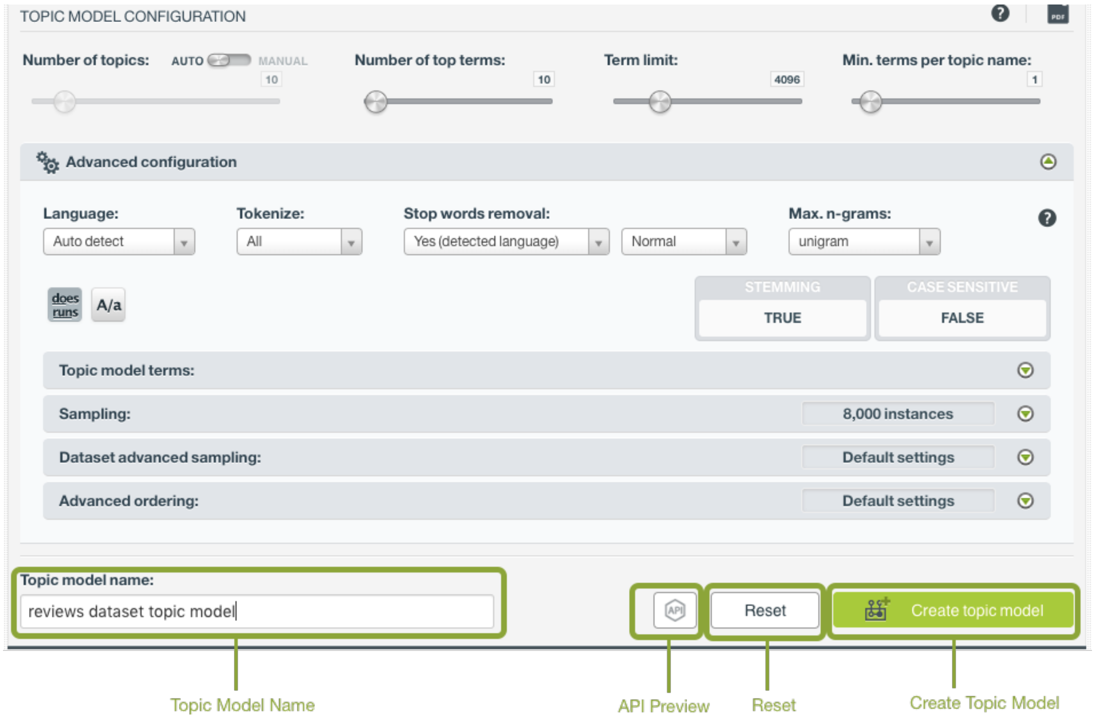
RANGE 1 - 1,259; SAMPLING Random; REPLACEMENT NO; OUT OF BAG NO

Topic model name: mental-health-in-tech-survey (1) dataset's topic model

Figure 4.13: Sampling options for topic models

4.7 Creating Topic Models with Configured Options

After finishing the configuration of your options, you can change the default topic model name in the editable text box. Then you can click on the Create topic model button to create the new topic model, or reset the configuration by clicking on the Reset button.



TOPIC MODEL CONFIGURATION

Number of topics: AUTO MANUAL 10

Number of top terms: 10

Term limit: 4096

Min. terms per topic name: 1

Advanced configuration

Language: Auto detect

Tokenize: All

Stop words removal: Yes (detected language) Normal

Max. n-grams: unigram

STEMMING TRUE

CASE SENSITIVE FALSE

Topic model terms:

Sampling: 8,000 instances

Dataset advanced sampling: Default settings

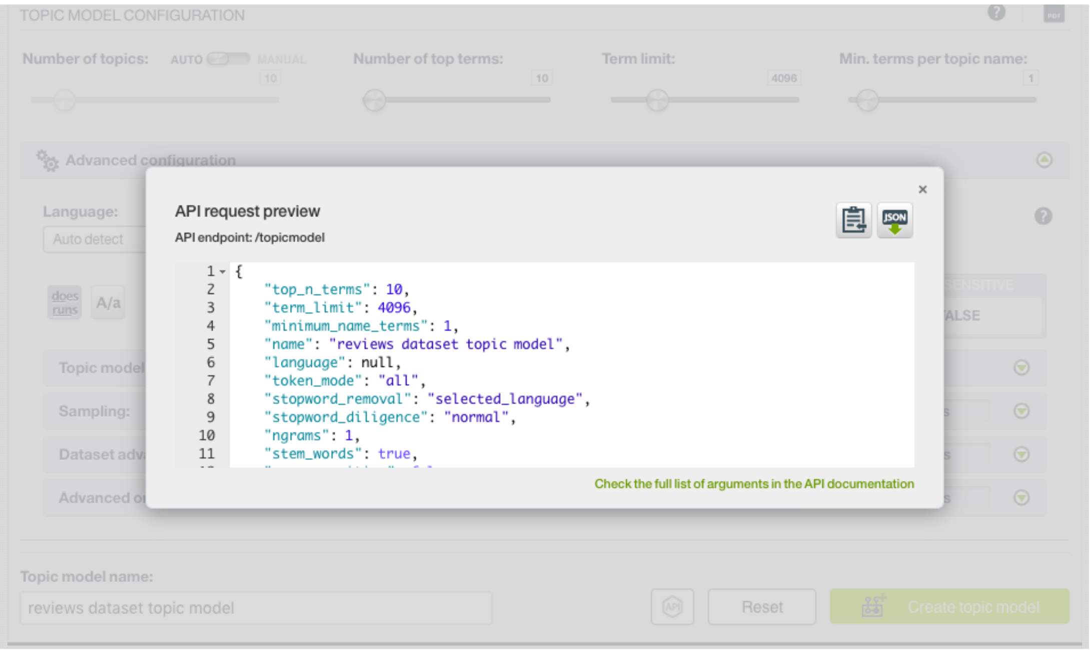
Advanced ordering: Default settings

Topic model name: reviews dataset topic model

Figure 4.14: Create topic model after configuration

4.8 API Request Preview

The API Request Preview button is in the middle on the bottom of the configuration panel, next to the Reset button (See (Figure 4.14)). This is to show how to create the topic model programmatically: the endpoint of the REST API call and the JSON that specifies the arguments configured in the panel. Please see (Figure 4.15) below:



API request preview

API endpoint: /topicmodel

```
{
    "top_n_terms": 10,
    "term_limit": 4096,
    "minimum_name_terms": 1,
    "name": "reviews dataset topic model",
    "language": null,
    "token_mode": "all",
    "stopword_removal": "selected_language",
    "stopword_diligence": "normal",
    "ngrams": 1,
    "stem_words": true,
```

Check the full list of arguments in the API documentation

Figure 4.15: Topic model API request preview

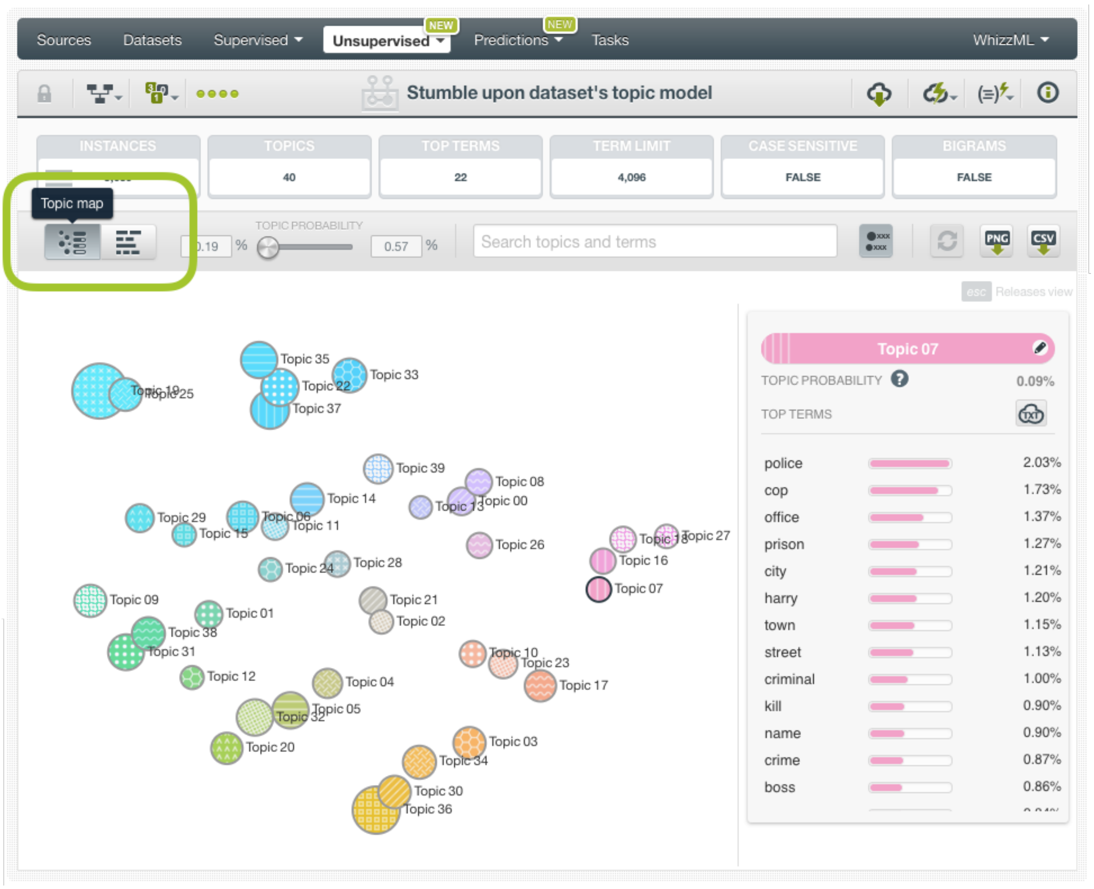
There are options on the upper right to either export the JSON or copy it to clipboard. On the bottom there is a link to the API documentation for topic models, in case you need to check any of the possible values or want to extend your knowledge in the use of the API to automate your workflows.

Please note: when a default value for an argument is used in the chosen configuration, the argument won't appear in the generated JSON. Because during API calls, default values are used when arguments are missing, there is no need to send them in the creation request.

CHAPTER 5

Visualizing Topic Models

BigML topic models are composed of two main views: the **Topic Map** and the **Term Chart**. You can switch them by clicking in the icons on the top menu. (See Figure 5.1.)



INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
[illegible]	40	22	4,096	FALSE	FALSE

TOPIC PROBABILITY 0.19 % – 0.57 %

Topic 07

TOPIC PROBABILITY 0.09%

TOP TERMS

police		2.03%
cop		1.73%
office		1.37%
prison		1.27%
city		1.21%
harry		1.20%
town		1.15%
street		1.13%
criminal		1.00%
kill		0.90%
name		0.90%
crime		0.87%
boss		0.86%

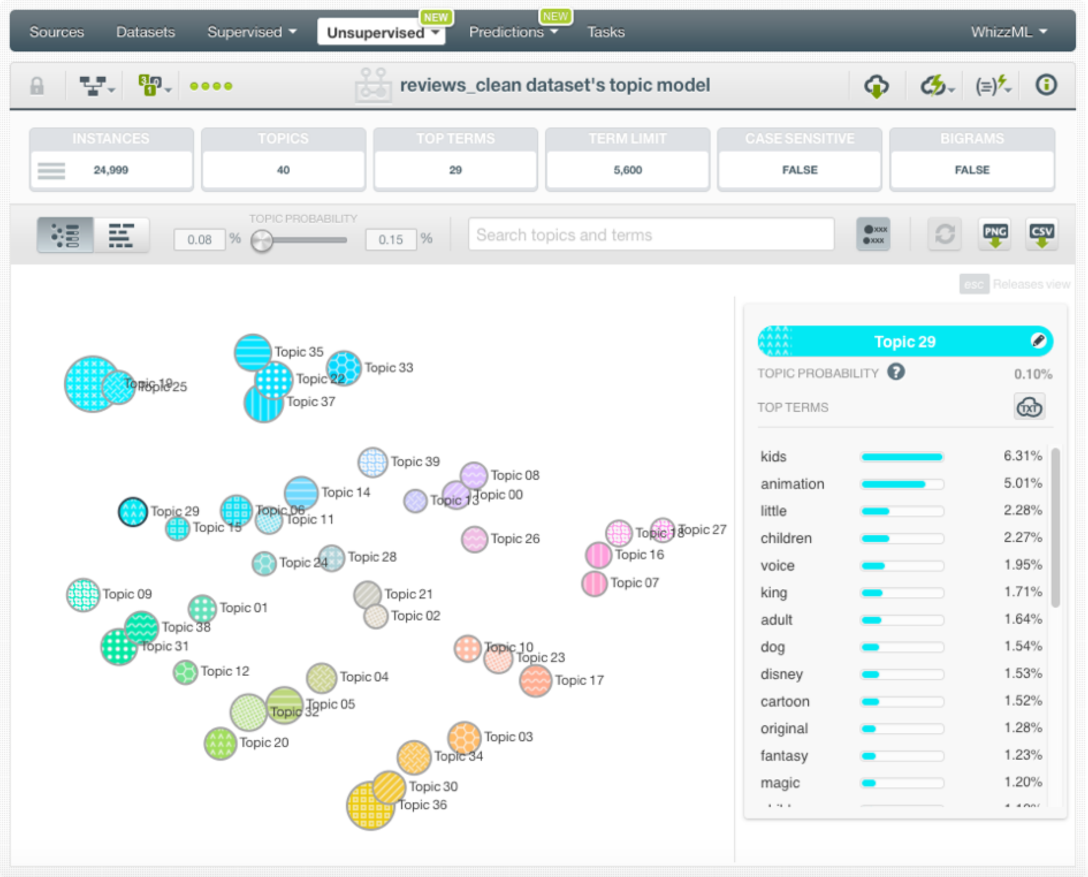
Figure 5.1: Switch views by clicking in the corresponding icons

5.1 Topic Map

The topic map is ideal to get an **overview** of both the **topic importances** in the dataset and the **relationship** between them. In this view you can see all the discovered topics mapped as circles. Each circle size depicts the topic importance in the dataset used to create the model. This importance is measured as the **topic probability** — this is the average probability of the topic to appear in a given instance of the original dataset. The distance between the circles represents the thematic closeness of topics, i.e., two topics that have terms in common and/or usually appear together in the same documents will be closer than topics which are not thematically related.

By mousing over each topic you can see the **top most important terms** for the topic. You can configure this number before creating the topic model using the option explain in Section 4.2. The top terms are sorted in ascending order by their probability in the topic. All term probabilities for a given topic should sum up 100%. A given term can be attributed to more than one topic, e.g., the word “bank” may be found in a topic related to finances, but also in a topic related to geology (river bank). Learn more about topics in Chapter 2.

Click on the topic circle or press Shift from your keyboard to freeze this view. (See Figure 5.2.) You can release it again by pressing Escape .



reviews_clean dataset's topic model

INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	29	5,600	FALSE	FALSE

TOPIC PROBABILITY 0.08 % 0.15 %

Topic 29

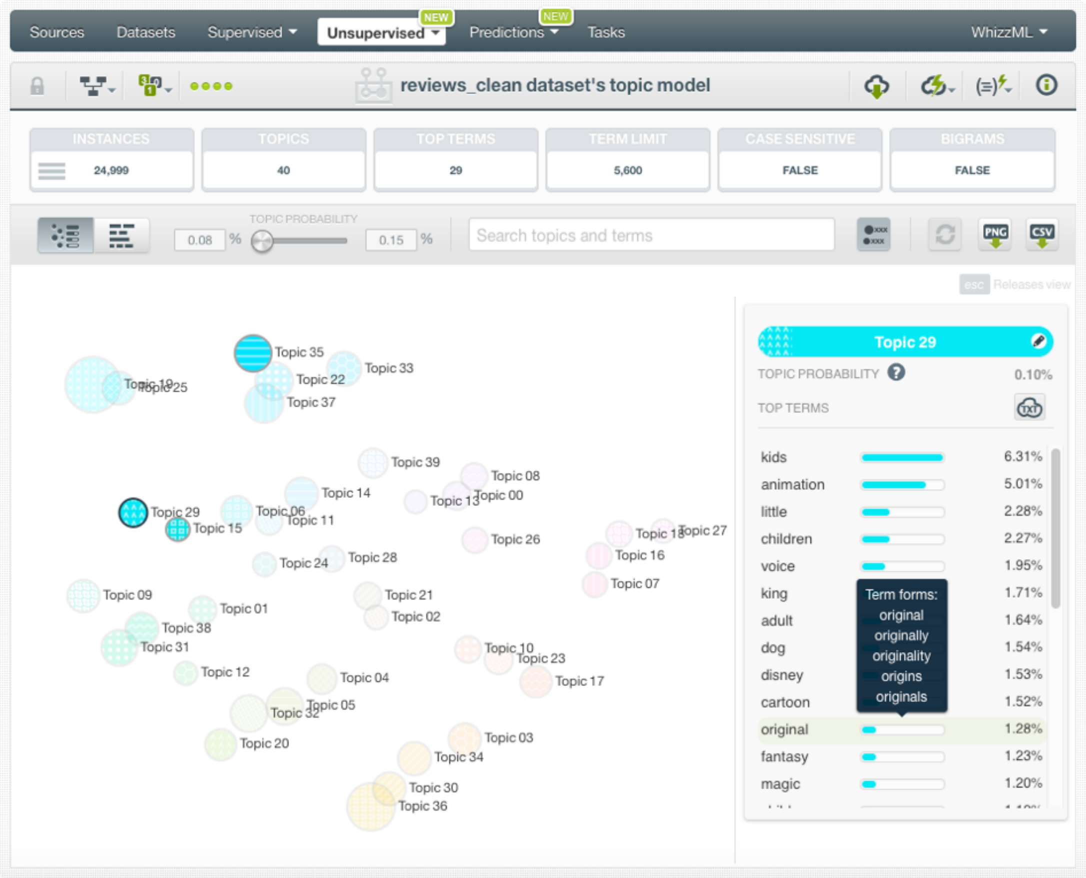
TOPIC PROBABILITY 0.10%

TOP TERMS

kids	6.31%
animation	5.01%
little	2.28%
children	2.27%
voice	1.95%
king	1.71%
adult	1.64%
dog	1.54%
disney	1.53%
cartoon	1.52%
original	1.28%
fantasy	1.23%
magic	1.20%

Figure 5.2: Topic map view

If you mouse over a **term** within a topic, the rest of the topics containing that term will be highlighted. Also a list containing all the **stemmed forms** for that term will appear. (See Figure 5.3.) Stemming is the process of taking just the lexeme for the terms, e.g. the terms “great”, “greatness” and “greater”, are considered the same term, since they all have the same lexeme: “great”. Please keep in mind that topic models always apply stemming to the model vocabulary. (See Chapter 2.)



INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	29	5,600	FALSE	FALSE

Topic 29

TOPIC PROBABILITY 0.10%

TOP TERMS

kids	6.31%
animation	5.01%
little	2.28%
children	2.27%
voice	1.95%
king	1.71%
adult	1.64%
dog	1.54%
disney	1.53%
cartoon	1.52%
original	1.28%
fantasy	1.23%
magic	1.20%

Term forms:
original
originally
originality
origins
originals

Figure 5.3: Stemmed forms for terms

BigML provides a set of filters and options for you to get a better visualization of your topics in the topic map:

- Filter topics by their probability using the **topic probability slider**. (See Figure 5.4.) As the number the number of instances and topics is higher, topic probabilities tend to be lower. It is not unusual to find topic probabilities below 1% (this is te average probability of the topic to appear in an instance along the entire dataset).

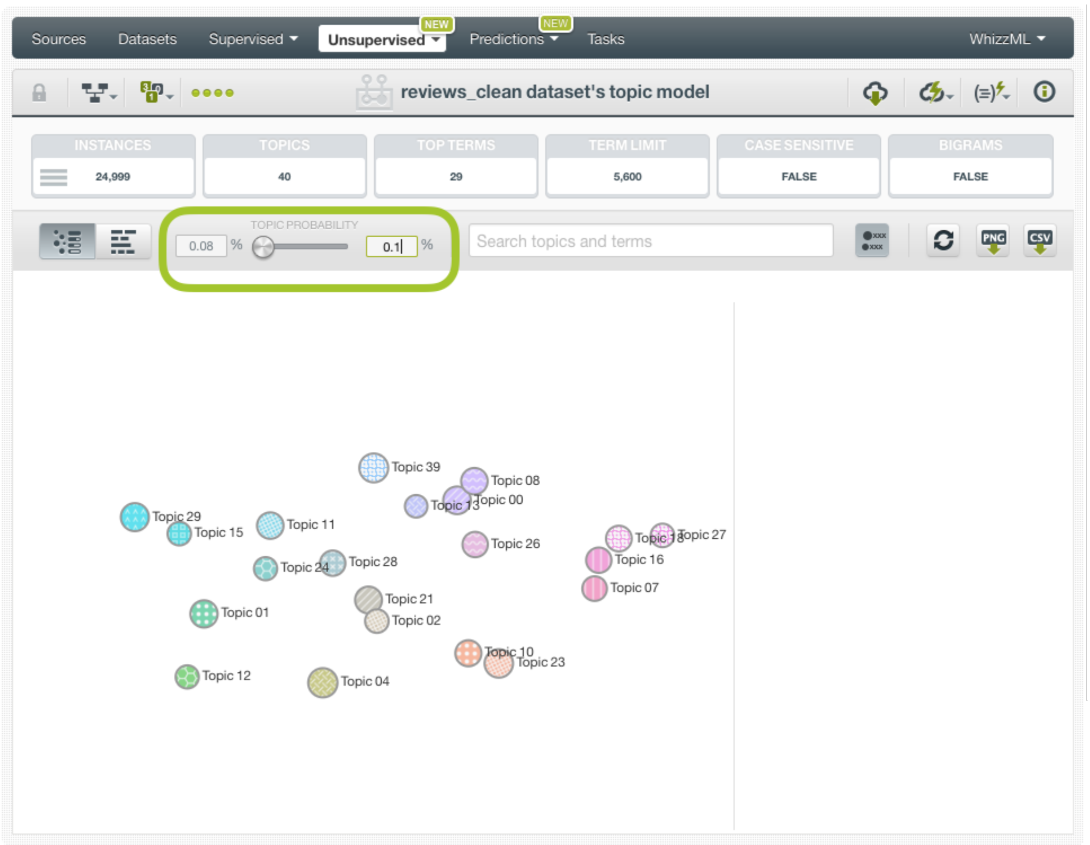
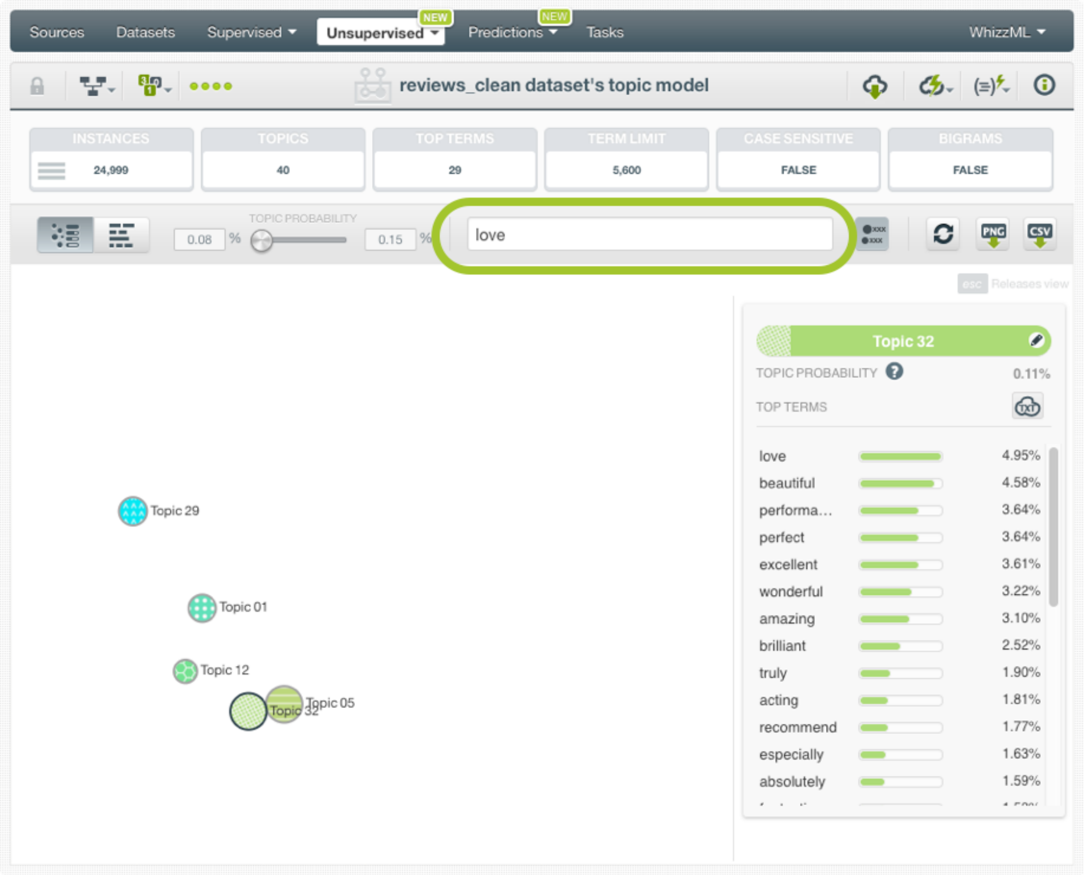


Figure 5.4: Filter by topic probabilities

- Search topics and terms using the **search box** on the top menu. If you type any topic name or term, the chart will show only the topics containing that information. (See Figure 5.5.)

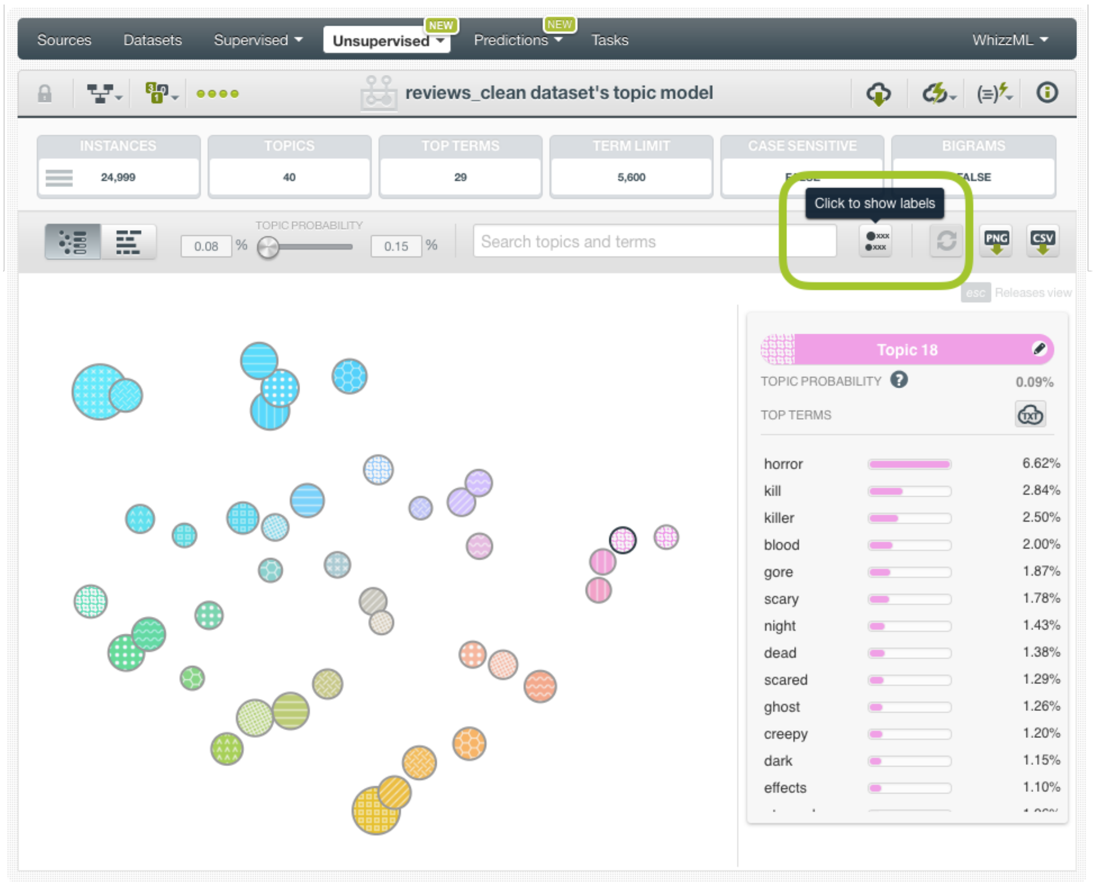


INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	29	5,600	FALSE	FALSE

love	4.95%
beautiful	4.58%
performa...	3.64%
perfect	3.64%
excellent	3.61%
wonderful	3.22%
amazing	3.10%
brilliant	2.52%
truly	1.90%
acting	1.81%
recommend	1.77%
especially	1.63%
absolutely	1.59%

Figure 5.5: Search terms and topics

- You can show or hide the labels for the **topic names** by using the option in the top menu. (See Figure 5.6.)



INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	29	5,600	FALSE	FALSE

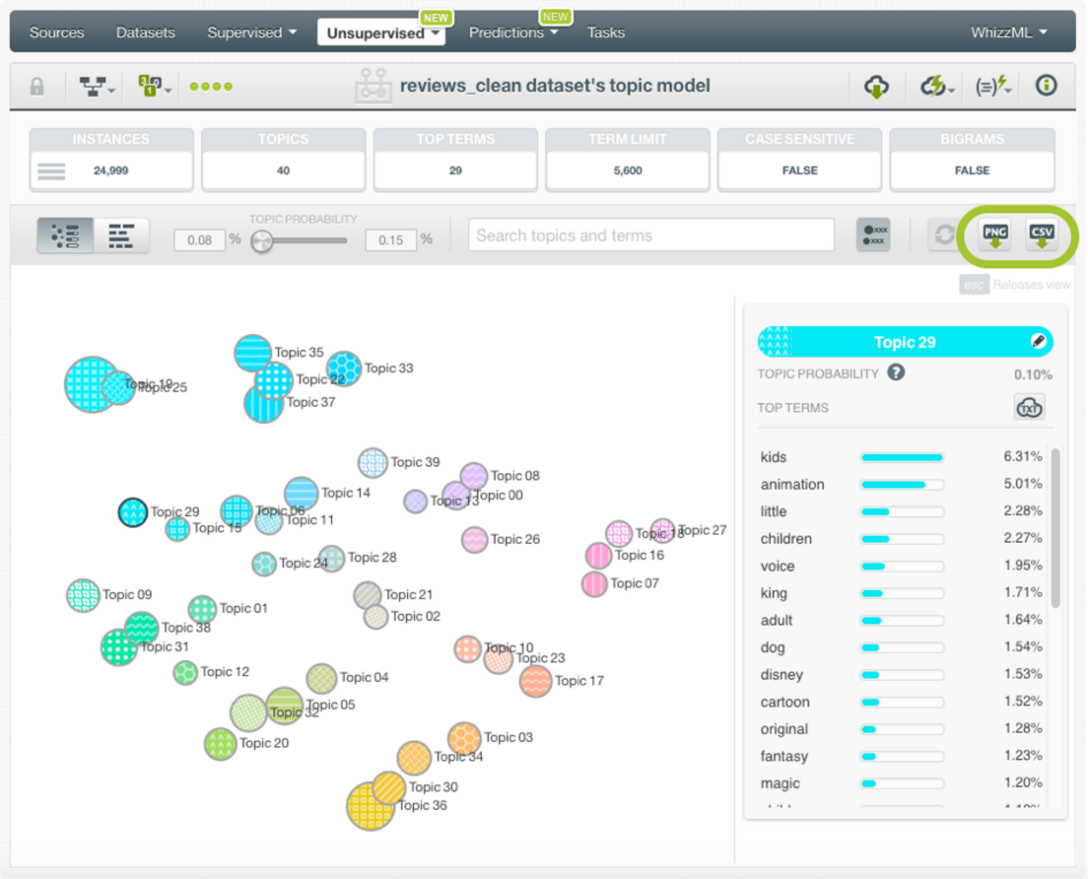
TOPIC PROBABILITY 0.09%

horror	6.62%
kill	2.84%
killer	2.50%
blood	2.00%
gore	1.87%
scary	1.78%
night	1.43%
dead	1.38%
scared	1.29%
ghost	1.26%
creepy	1.20%
dark	1.15%
effects	1.10%

Figure 5.6: Show or hide topic names

Moreover, you can export the chart as an image and the model in CSV format (see Figure 5.7.):

- **Export chart in PNG**: you can export the topic map as an image.
- **Export the model in CSV**: you can export the model in a CSV file. The rows in the CSV file will be the terms of the model and the columns, the topics. Each term will have a set of probabilities associated, one by topic.



INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	29	5,600	FALSE	FALSE

TOPIC PROBABILITY 0.08 % – 0.15 %

Topic 29

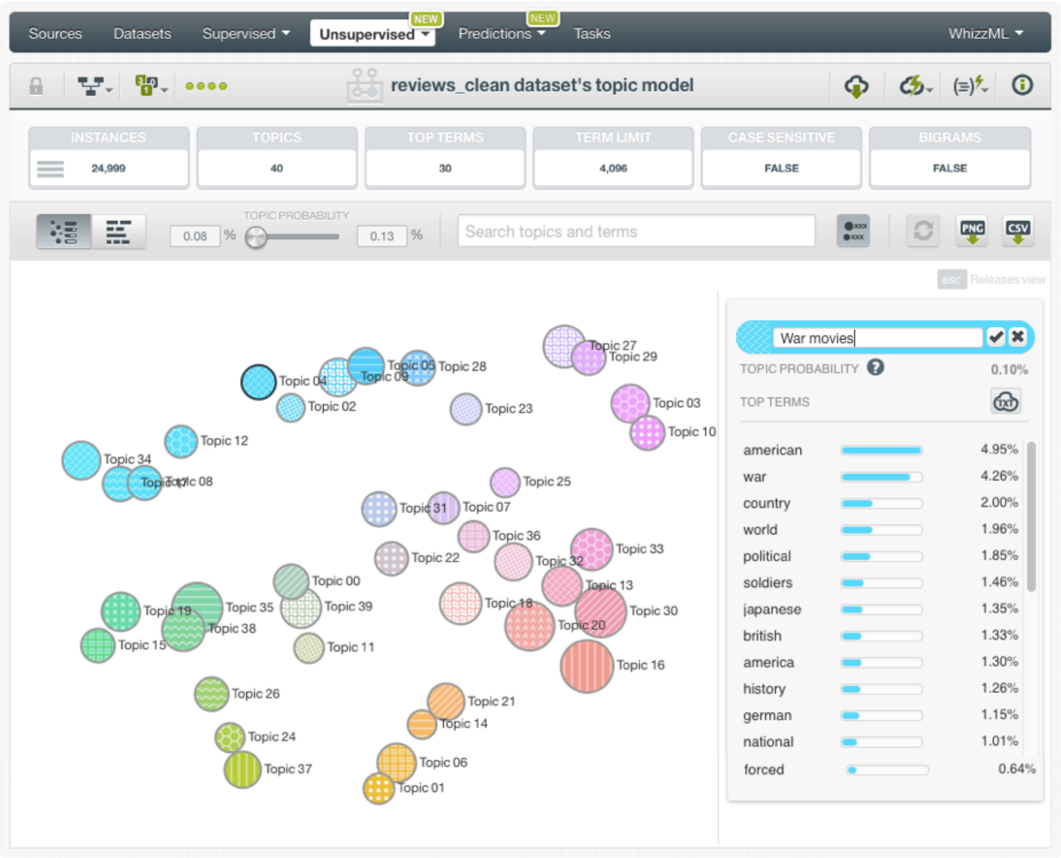
TOPIC PROBABILITY 0.10%

TOP TERMS

kids	6.31%
animation	5.01%
little	2.28%
children	2.27%
voice	1.95%
king	1.71%
adult	1.64%
dog	1.54%
disney	1.53%
cartoon	1.52%
original	1.28%
fantasy	1.23%
magic	1.20%

Figure 5.7: Topic map export options

Most topic model images in this document show generic enumerated **names for the topics** starting at “Topic 00”, “Topic 01” and so on. This is because we configured our topic model before creating it (see Section 4.4), otherwise, BigML chooses the top term of each topic to set the names by default. You can **edit** the topic names by clicking in the edition icon shown in Figure 5.8. Editing the topic names to set a proper name related to the topic content is very useful for inspecting your predictions afterwards, because you will be able to understand the most relevant topics for a given instance without having to look at its terms.



INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	30	4,096	FALSE	FALSE

War movies

TOPIC PROBABILITY 0.10%

TOP TERMS

american	4.95%
war	4.26%
country	2.00%
world	1.96%
political	1.85%
soldiers	1.46%
japanese	1.35%
british	1.33%
america	1.30%
history	1.26%
german	1.15%
national	1.01%
forced	0.64%

Figure 5.8: Edit topic names

You can also see the top terms in a **tag cloud** to get a more insightful view of your term importances in a given topic. You can download the tag cloud in SVG and PNG formats.(See Figure 5.9.)

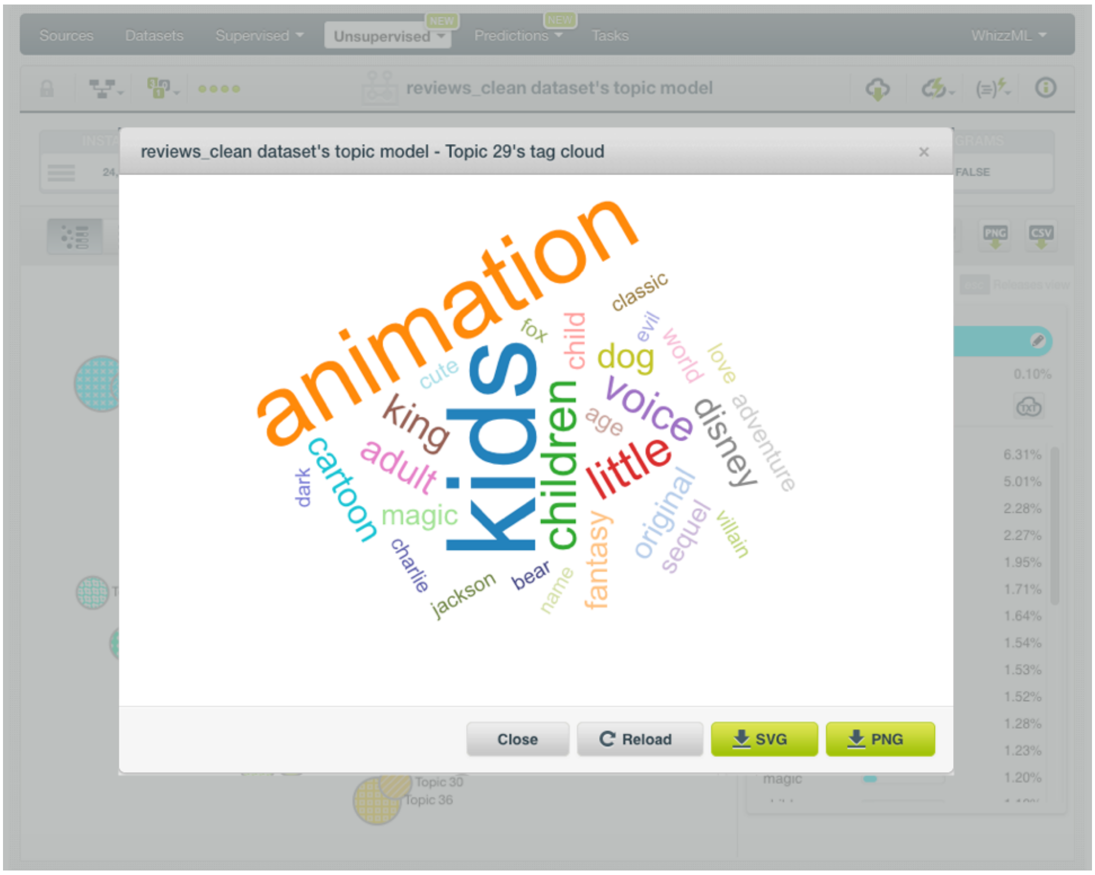


Figure 5.9: Topic terms in tag cloud

5.2 Term Chart

The term chart is probably the most appropriate way to get an **overview of your topic themes**. In this view you will find the topics and their terms in a bar chart. The topics are sorted by name on the vertical axis, if you mouse over the topic name, a tooltip with the topic probability in the dataset will appear. The terms are plotted along the horizontal axis in a bar chart where the size of the bar represents the term probabilities for a given topic. By default, BigML shows eight terms per topic but you can select up to 15 terms.

Depending on your topic model, you may need to make use of the **filters** and other **chart options** in order to correctly visualize your topics:

- Filter terms by their probability using the **term probability slider**. This filter is really useful when your topics contain terms with very high probabilities, since filtering them will allow you to better visualize the rest of the terms. You can see an example in Figure 5.10. The left-hand-side image contains two terms with a very high probability (the long blue bars) so they overshadow the other terms. If you filter them using the term probability filter, the rest of the terms are better visualized.

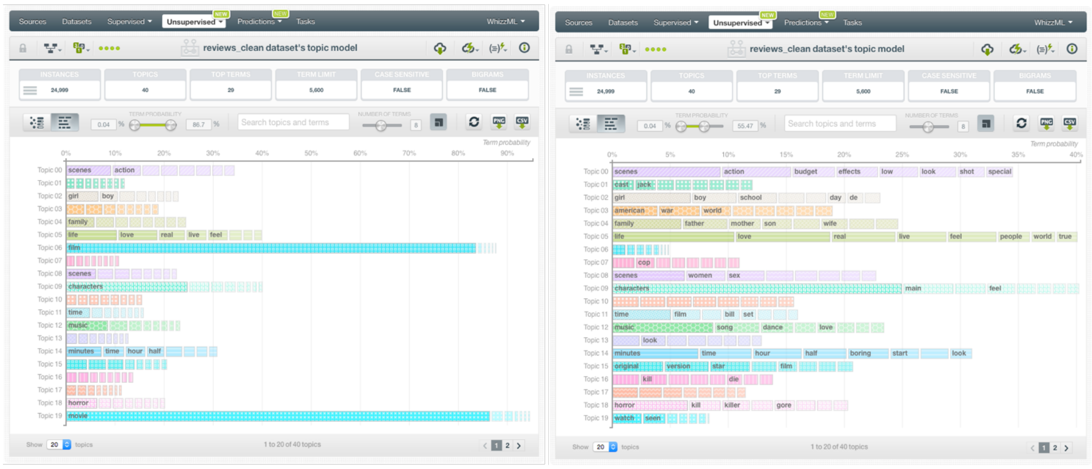


Figure 5.10: Term probability filter

- Search topics and terms using the **search box** on the top menu. If you type any topic name or term, the chart will show only the topics containing that information. (See Figure 5.11.)

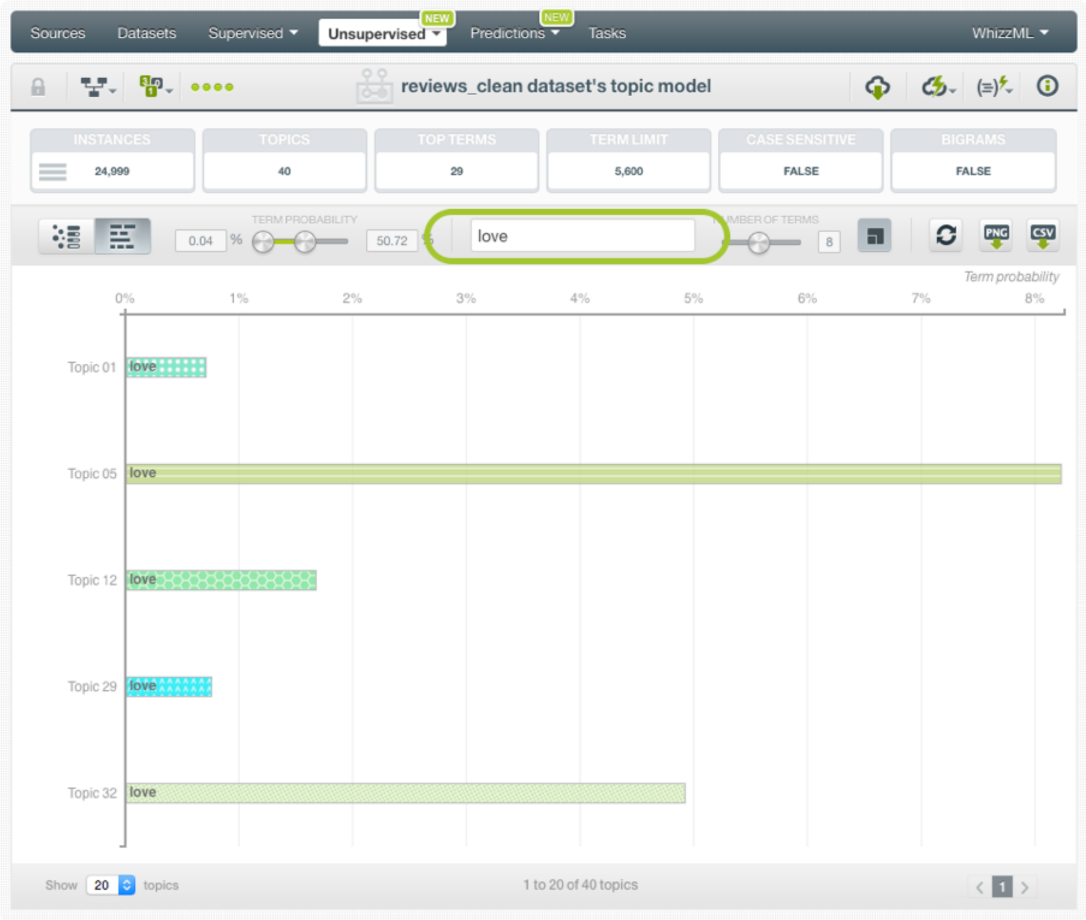
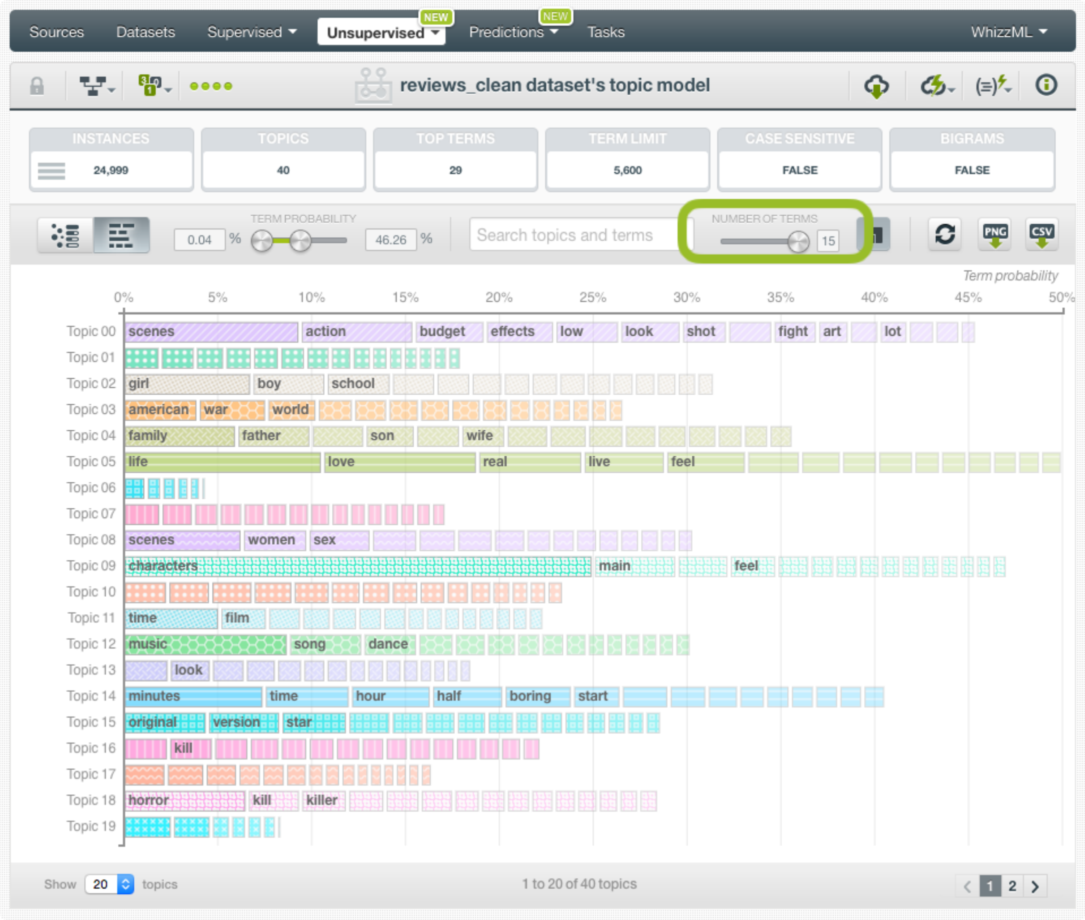


Figure 5.11: Search terms and topics

- Select the maximum terms shown per topic in the chart using the **number of terms** slider. You can select up to 15 terms per topic. Terms are always ordered by probability from left to right. (See Figure 5.12.)



INSTANCES	TOPICS	TOP TERMS	TERM LIMIT	CASE SENSITIVE	BIGRAMS
24,999	40	29	5,600	FALSE	FALSE

TERM PROBABILITY 0.04 % 46.26 %

Search topics and terms

NUMBER OF TERMS 15

Figure 5.12: Select the maximum number of top terms to display in the chart

- The maximum value of the horizontal axis is set according to the highest sum of term probabilities for the topic model. You can adjust the axis scale so you its maximum value is **dynamically adjusted** to the maximum sum of term probabilites shown at any time in the chart. Therefore you can better visualize the terms for each topic. (See Figure 5.13.)

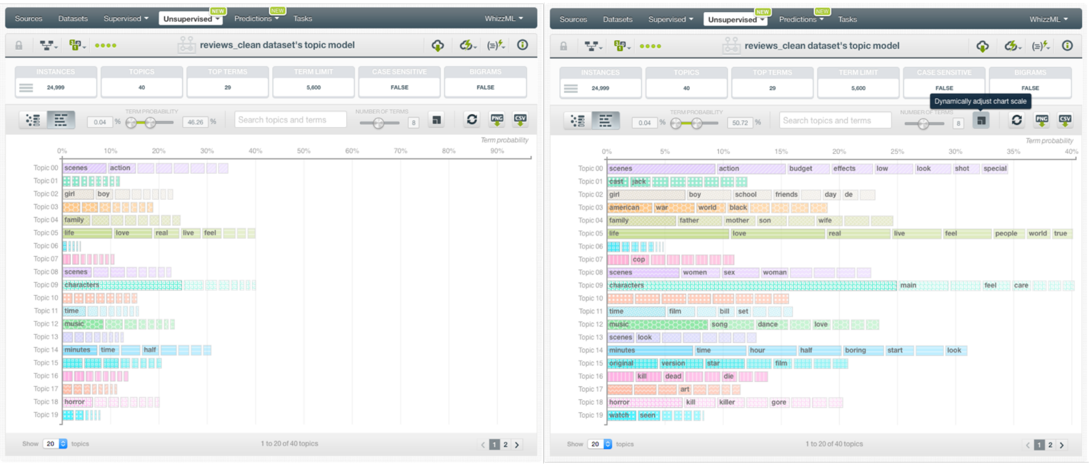
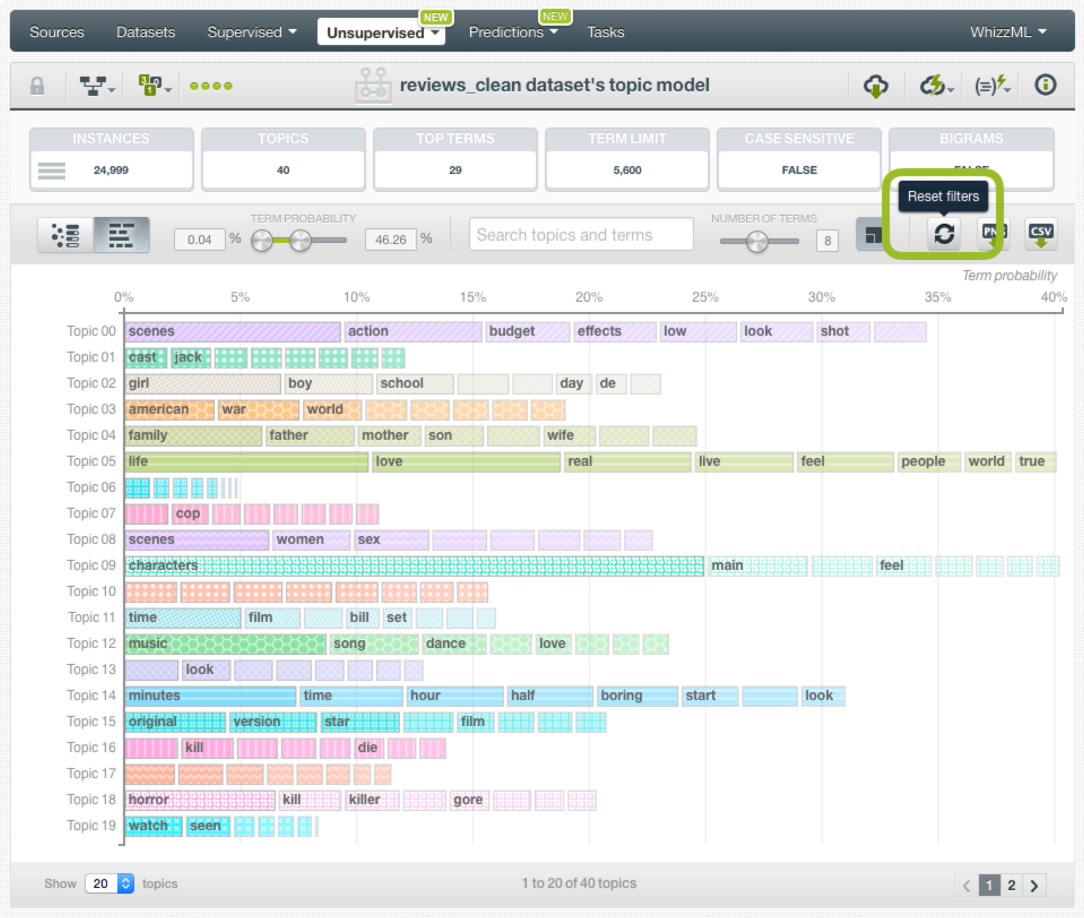

reviews_clean dataset's topic model
Dynamically adjust chart scale
Term probability
Topic 00
Topic 19
1 to 20 of 40 topics

Figure 5.13: Adjust the axis scale to current filters

- You can reset all your filters to their default values by clicking the **reset filters** option. (See Figure 5.14.)


Sources
Datasets
Supervised
NEW
Unsupervised
NEW
Predictions
Tasks
WhizzML
reviews_clean dataset's topic model
Reset filters
Search topics and terms
Term probability
0%
5%
10%
15%
20%
25%
30%
35%
40%
Topic 00 scenes action budget effects low look shot
Topic 01 cast jack
Topic 02 girl boy school day de
Topic 03 american war world
Topic 04 family father mother son wife
Topic 05 life love real live feel people world true
Topic 06
Topic 07 cop
Topic 08 scenes women sex
Topic 09 characters main feel
Topic 10
Topic 11 time film bill set
Topic 12 music song dance love
Topic 13 look
Topic 14 minutes time hour half boring start look
Topic 15 original version star film
Topic 16 kill die
Topic 17
Topic 18 horror kill killer gore
Topic 19 watch seen
1
2

INSTANCES 24,999
TOPICS 40
TOP TERMS 29
TERM LIMIT 5,600
CASE SENSITIVE FALSE
BIGRAMS FALSE

TERM PROBABILITY 0.04 % 46.26 %
NUMBER OF TERMS 8

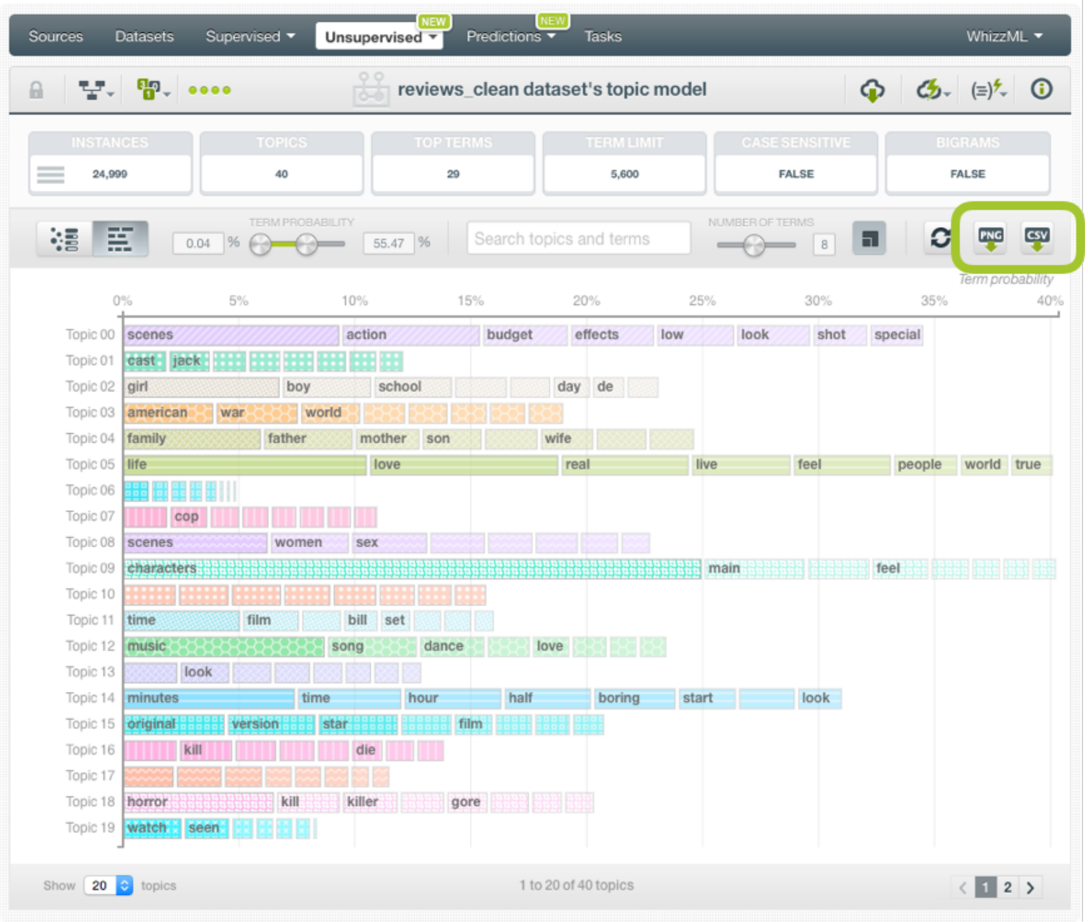
Show 20 topics

1 to 20 of 40 topics

Figure 5.14: Reset filters option

You can export the term chart as an image and the model in CSV format:

- **Export chart in PNG**: you can export the term chart as an image including or excluding the legends. The legend includes the list of topics along with a list of terms by topic.
- **Export the model in CSV**: explained in Section 5.1.



reviews_clean dataset's topic model

INSTANCES 24,999

TOPICS 40

TOP TERMS 29

TERM LIMIT 5,600

CASE SENSITIVE FALSE

BIGRAMS FALSE

TERM PROBABILITY 0.04 % 55.47 %

NUMBER OF TERMS 8

Show 20 topics

1 to 20 of 40 topics

Sources
Datasets
Supervised
Unsupervised
NEW
Predictions
NEW
Tasks
WhizzML
Search topics and terms
PNG
CSV
Term probability
0%
5%
10%
15%
20%
25%
30%
35%
40%
Topic 00 scenes action budget effects low look shot special
Topic 01 cast jack
Topic 02 girl boy school day de
Topic 03 american war world
Topic 04 family father mother son wife
Topic 05 life love real live feel people world true
Topic 06
Topic 07 cop
Topic 08 scenes women sex
Topic 09 characters main feel
Topic 10
Topic 11 time film bill set
Topic 12 music song dance love
Topic 13 look
Topic 14 minutes time hour half boring start look
Topic 15 original version star film
Topic 16 kill die
Topic 17
Topic 18 horror kill killer gore
Topic 19 watch seen
1
2

Figure 5.15: Term chart export options

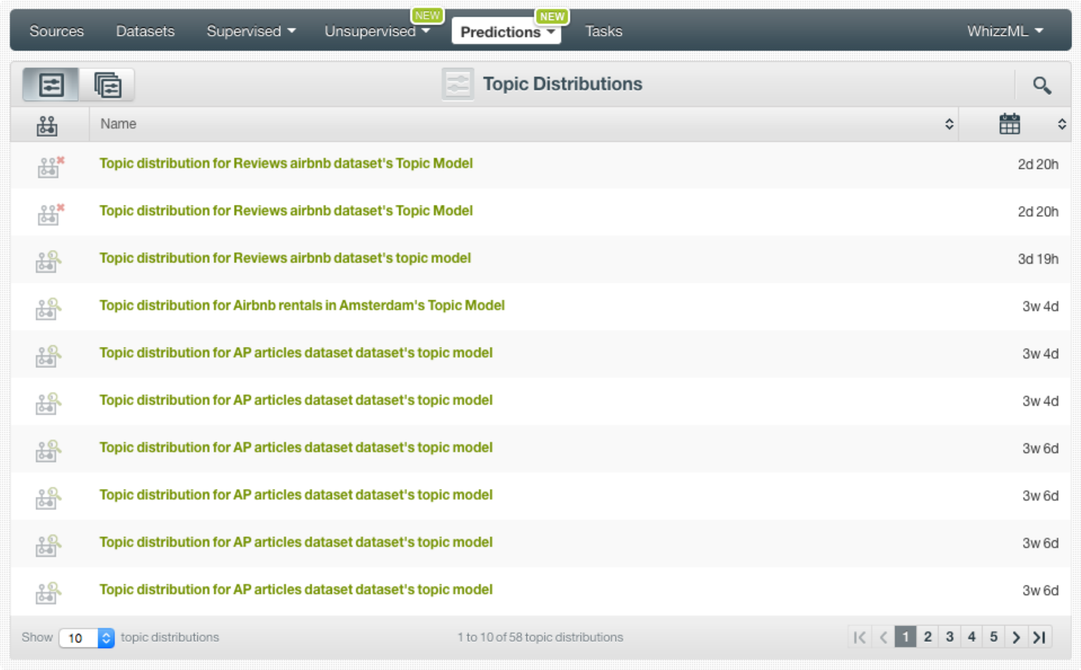
CHAPTER 6

Topic Model Predictions: Topic Distributions

6.1 Introduction

The main goal of building a topic model is to find the relevant topics in your dataset. The calculation of the topic probabilities over your text instances is referred to as **topic distributions** in BigML. You can create topic distributions either for **single instances**, i.e., one by one, or for **multiple instances** simultaneously, i.e., in batch. Each instance will have a set of probabilities associated, one per topic, indicating the relevance of that topic for the instance. The sum of all topic probabilities for a given instance must be 100%.

The predictions tab in the main menu of the BigML Dashboard is where all of your saved topic distributions are listed (Figure 6.1). In the topic distribution list view, you can see the icon for the **topic model** used for each topic distribution, the **Name** and the **Age** (time since the topic distribution was created). You can also search your topic distributions by name clicking in the search menu option on the top right menu.



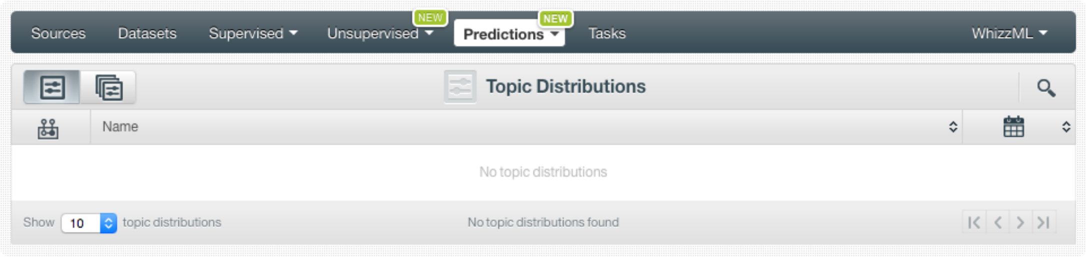
Name	
Topic distribution for Reviews airbnb dataset's Topic Model	2d 20h
Topic distribution for Reviews airbnb dataset's Topic Model	2d 20h
Topic distribution for Reviews airbnb dataset's topic model	3d 19h
Topic distribution for Airbnb rentals in Amsterdam's Topic Model	3w 4d
Topic distribution for AP articles dataset dataset's topic model	3w 4d
Topic distribution for AP articles dataset dataset's topic model	3w 4d
Topic distribution for AP articles dataset dataset's topic model	3w 6d
Topic distribution for AP articles dataset dataset's topic model	3w 6d
Topic distribution for AP articles dataset dataset's topic model	3w 6d
Topic distribution for AP articles dataset dataset's topic model	3w 6d

Show 10 topic distributions

1 to 10 of 58 topic distributions

Figure 6.1: Topic Distribution list view

By default, when you first create an account at BigML, or every time that you start a new project, your list view for predictions will be empty. (See Figure 6.2.)

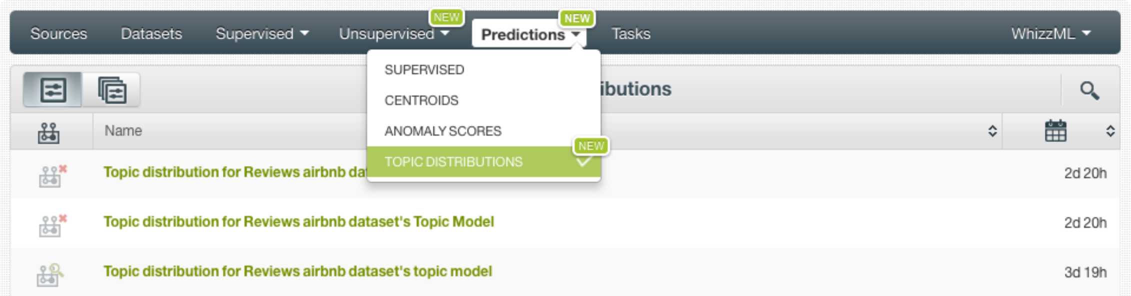


No topic distributions

No topic distributions found

Figure 6.2: Empty Dashboard topic distributions view

Topic distributions are saved under the prediction menu as shown in (Figure 6.3.)



Name	
Topic distribution for Reviews airbnb da	2d 20h
Topic distribution for Reviews airbnb dataset's Topic Model	2d 20h
Topic distribution for Reviews airbnb dataset's topic model	3d 19h

Figure 6.3: Topic distributions in prediction menu

From the topic distribution main view, you can select the list of your **topic distributions** or your **batch topic distributions** by clicking on the corresponding icons (see Figure 6.4 and Figure 6.5.)

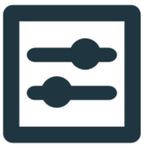


Figure 6.4: Single topic distribution icon

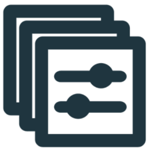


Figure 6.5: Batch topic distribution icon

6.2 Creating Topic Distributions

BigML provides two different ways to predict topic probabilities for new instances using your topic model:

- TOPIC DISTRIBUTION: to predict single instances.
- BATCH TOPIC DISTRIBUTION: to predict multiple instances simultaneously.

6.2.1 Topic Distribution

To predict the topic distributions for single instances BigML provides a form containing the fields used by the topic model so you can easily insert the input text and get an immediate response.

Follow these steps to create your topic distribution:

1. Click the TOPIC DISTRIBUTION option in the **1-click action menu**. (See Figure 6.6.)

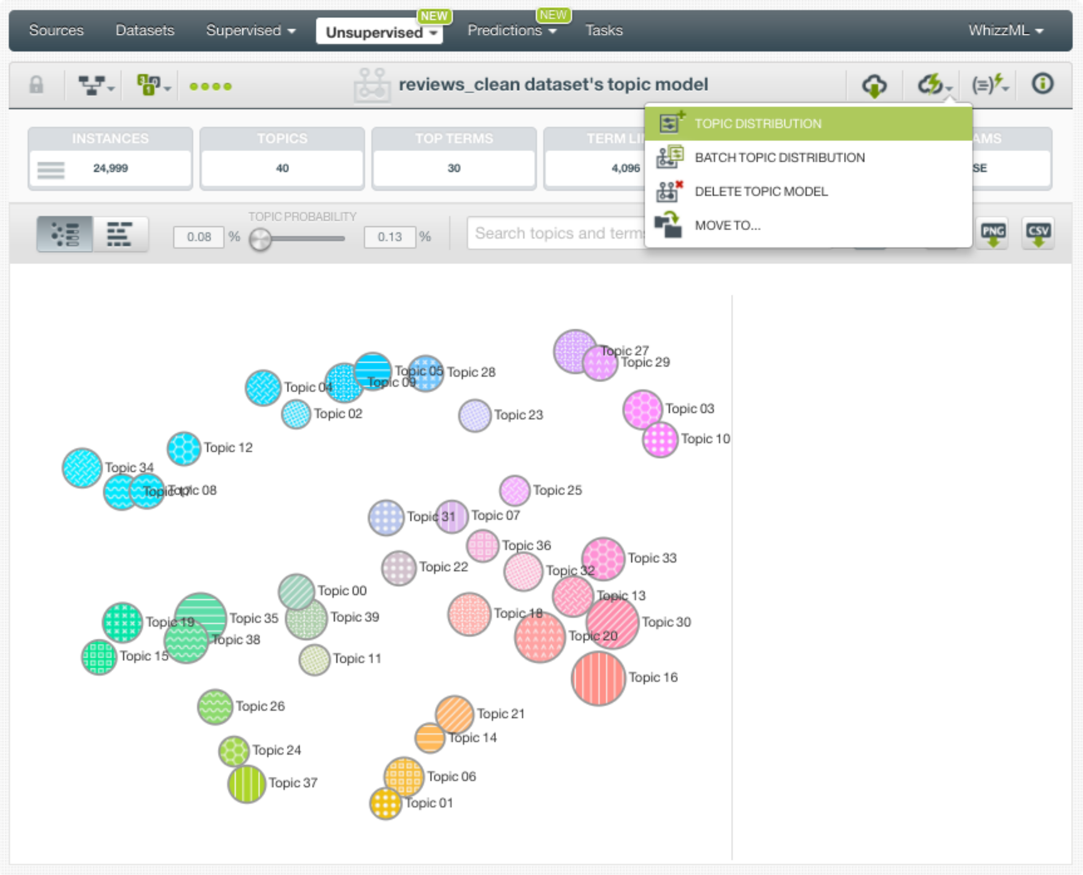
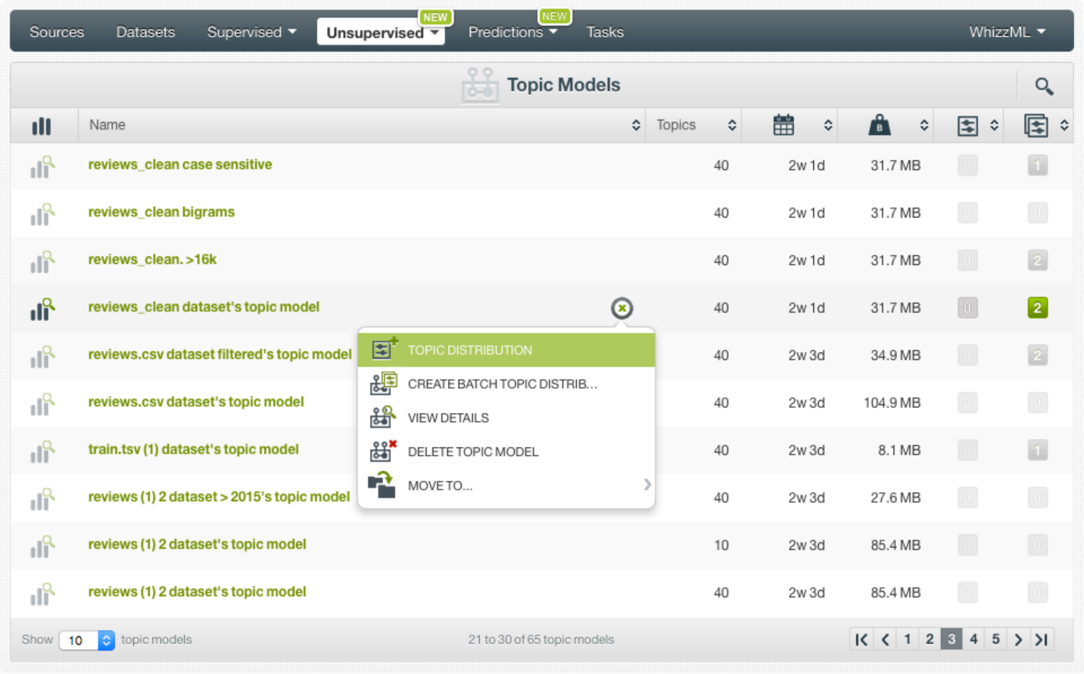


Figure 6.6: Predict option from topic model 1-click menu

Alternatively, click TOPIC DISTRIBUTION in the **pop up menu** from the topic model list view as shown in Figure 6.7.



Sources Datasets Supervised ▾ Unsupervised ▾ NEW Predictions ▾ NEW Tasks WhizzML ▾

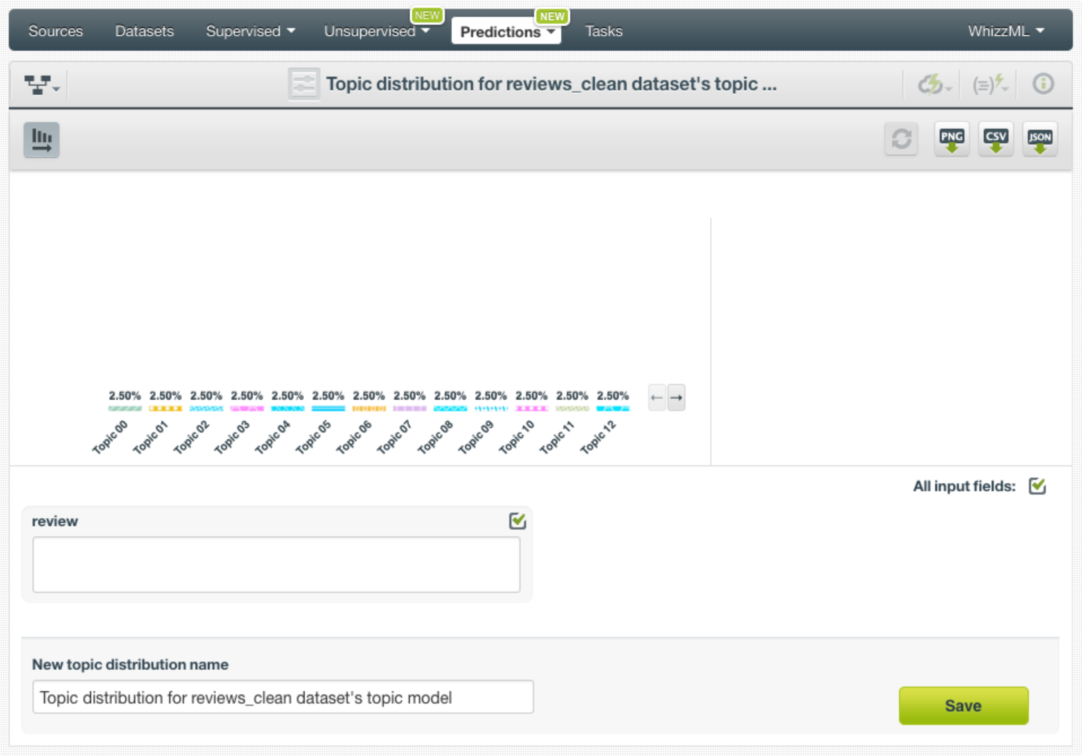
Topic Models

Name	Topics				
reviews_clean case sensitive	40	2w 1d	31.7 MB	0	1
reviews_clean bigrams	40	2w 1d	31.7 MB	0	0
reviews_clean. >16k	40	2w 1d	31.7 MB	0	2
reviews_clean dataset's topic model	40	2w 1d	31.7 MB	0	2
reviews.csv dataset filtered's topic model	40	2w 3d	34.9 MB	0	2
reviews.csv dataset's topic model	40	2w 3d	104.9 MB	0	0
train.tsv (1) dataset's topic model	40	2w 3d	8.1 MB	0	1
reviews (1) 2 dataset > 2015's topic model	40	2w 3d	27.6 MB	0	0
reviews (1) 2 dataset's topic model	10	2w 3d	85.4 MB	0	0
reviews (1) 2 dataset's topic model	40	2w 3d	85.4 MB	0	0

Show 10 topic models 21 to 30 of 65 topic models 1 2 3 4 5

Figure 6.7: Predict option from topic model pop up menu

2. You will be redirected to the **prediction form**, where you will get all the input fields used by the topic model to find the topics. (See Figure 6.8.)



Sources Datasets Supervised ▾ Unsupervised ▾ NEW Predictions ▾ NEW Tasks WhizzML ▾

Topic distribution for reviews_clean dataset's topic ...

All input fields: ☑

review ☑

New topic distribution name

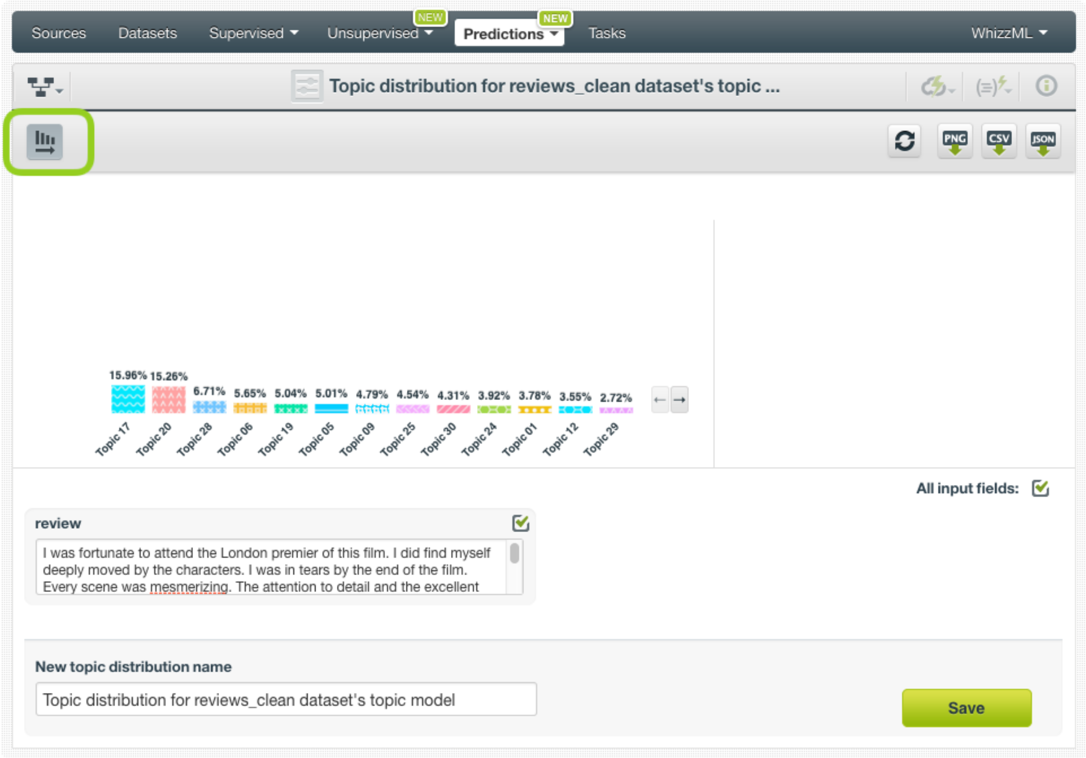
Topic distribution for reviews_clean dataset's topic model

Save

Figure 6.8: Topic distributions prediction form

3. **Select** the input fields and insert the **text** for which you want to obtain the topic distributions. You will instantly get the topic probability histogram for your input text. The topics are ordered by probability from left to right. If you want to sort them by name, click in the highlighted icon in

Figure 6.9.



Topic distribution for reviews_clean dataset's topic ...

All input fields: ☑

review ☑

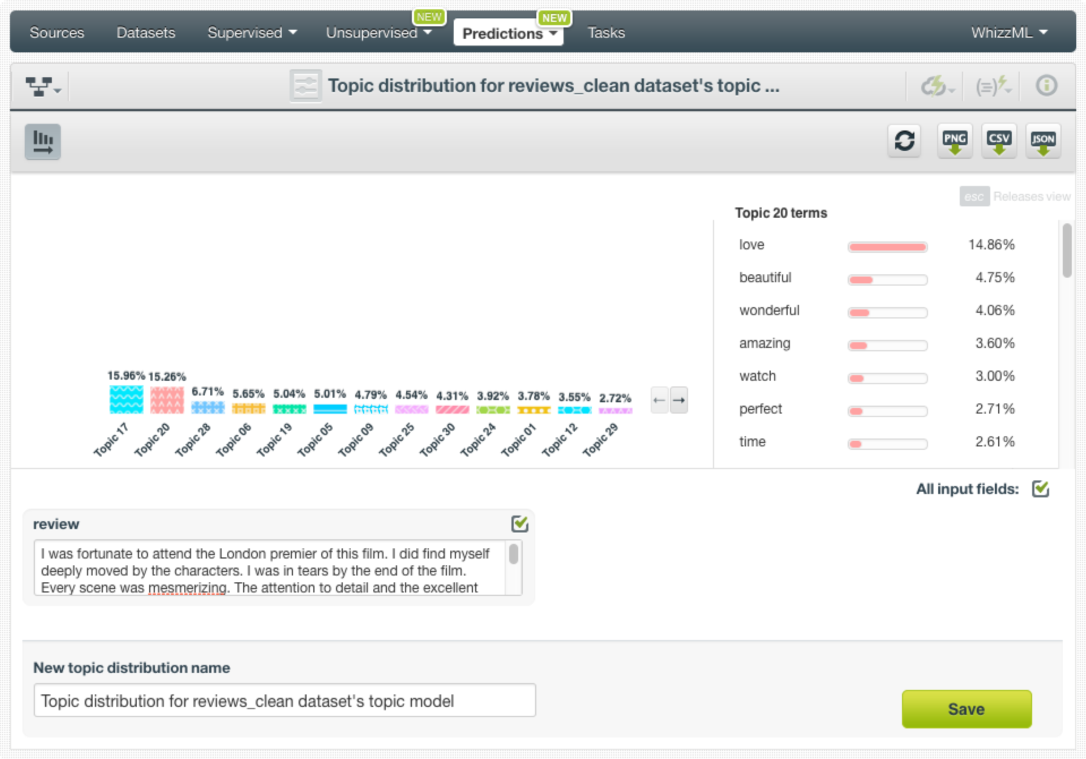
I was fortunate to attend the London premier of this film. I did find myself deeply moved by the characters. I was in tears by the end of the film. Every scene was mesmerizing. The attention to detail and the excellent

New topic distribution name

Topic distribution for reviews_clean dataset's topic model

Figure 6.9: Topic distribution histogram

4. By mousing over each topic bar you will see the top terms of the topic along with their probabilities to the right. (See Figure 6.10.)



Topic distribution for reviews_clean dataset's topic ...

Topic 20 terms

love	14.86%
beautiful	4.75%
wonderful	4.06%
amazing	3.60%
watch	3.00%
perfect	2.71%
time	2.61%

All input fields: ☑

review ☑

I was fortunate to attend the London premier of this film. I did find myself deeply moved by the characters. I was in tears by the end of the film. Every scene was mesmerizing. The attention to detail and the excellent

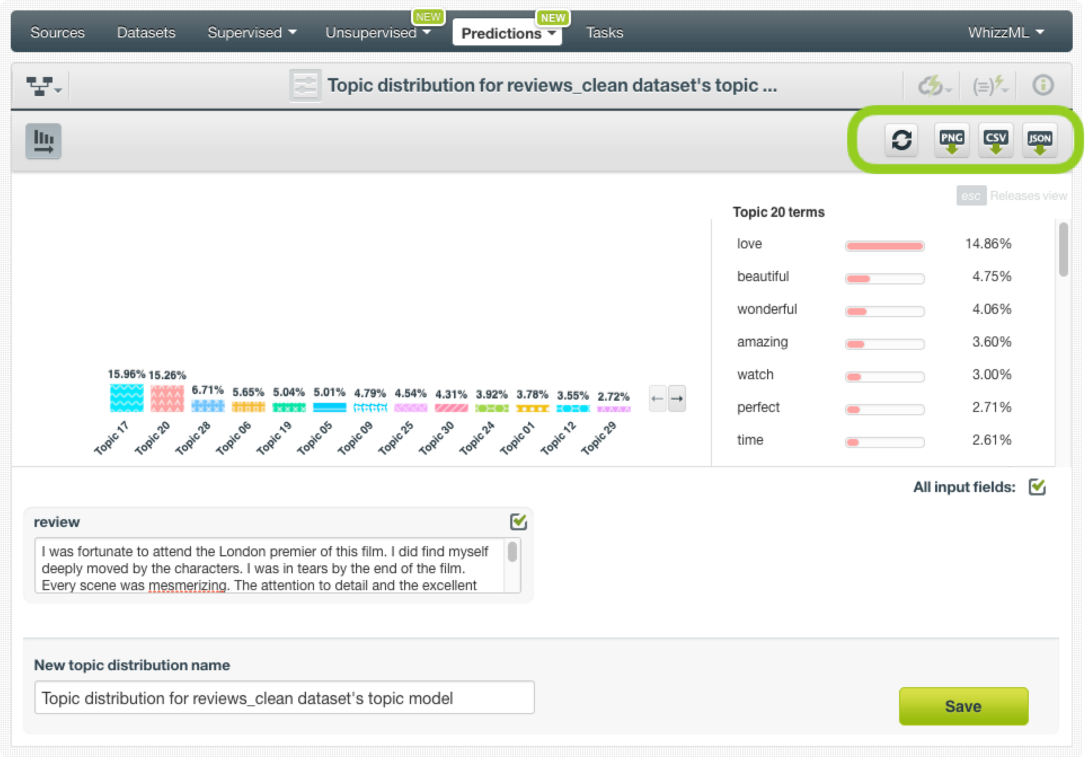
New topic distribution name

Topic distribution for reviews_clean dataset's topic model

Figure 6.10: Topic top term probabilities

5. Click the save button and your topic distribution will be saved in the prediction list view. (See Figure 6.1.)

6. Reset the input field values (to the last saved prediction) or download the topic distribution in PNG, CSV and JSON format by clicking in the corresponding icons shown in Figure 6.11



Topic 20 terms

love	14.86%
beautiful	4.75%
wonderful	4.06%
amazing	3.60%
watch	3.00%
perfect	2.71%
time	2.61%

All input fields: ☑

review ☑

I was fortunate to attend the London premier of this film. I did find myself deeply moved by the characters. I was in tears by the end of the film. Every scene was mesmerizing. The attention to detail and the excellent

New topic distribution name

Topic distribution for reviews_clean dataset's topic model

Figure 6.11: Topic distribution reset values and downloading options

6.2.2 Batch Topic Distribution

BigML batch topic distributions allow you to make predictions for multiple instances simultaneously. All you need is the **topic model** you want to use and a **dataset** containing the instances for which you want to obtain the topic distributions. You can use the same instances used to create the topic model or new instances. BigML will compute a set of probabilities, one per topic, for each instance.

Follow these steps to create a batch topic distribution:

1. Click BATCH TOPIC DISTRIBUTION option in the topic model **1-click action menu**. (See Figure 6.12.)

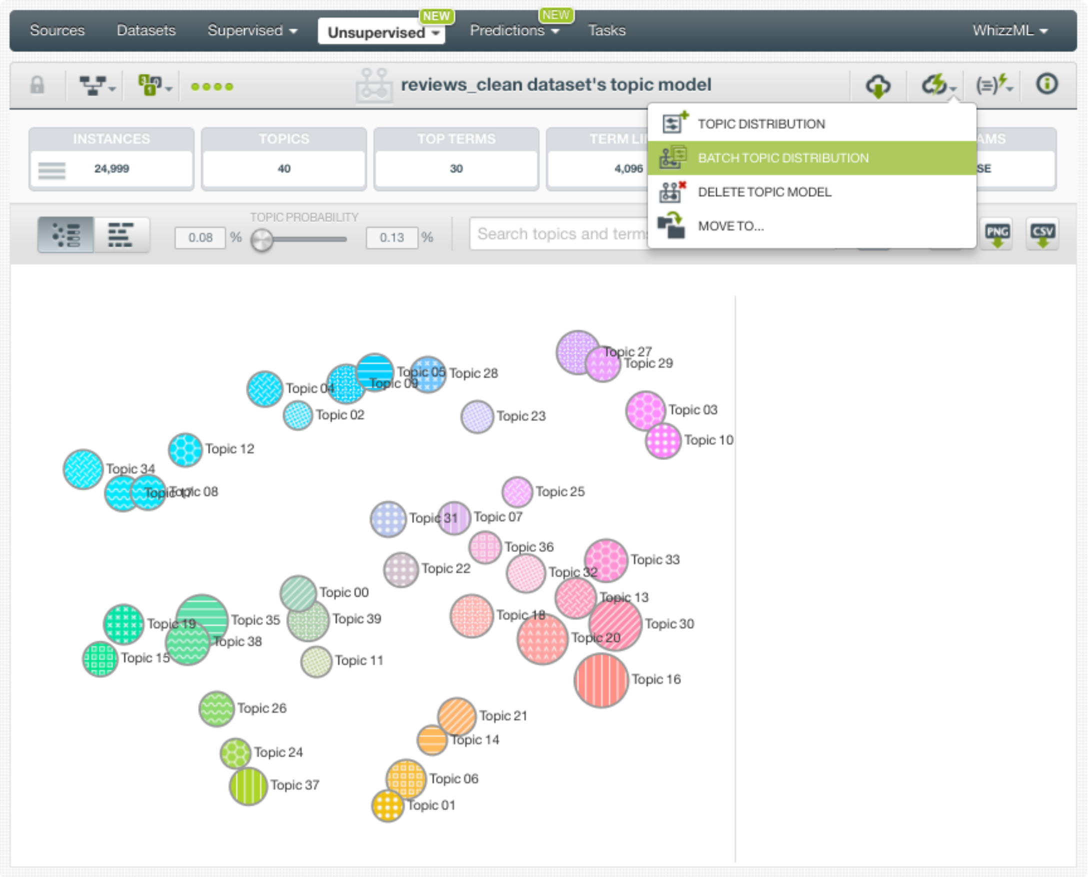
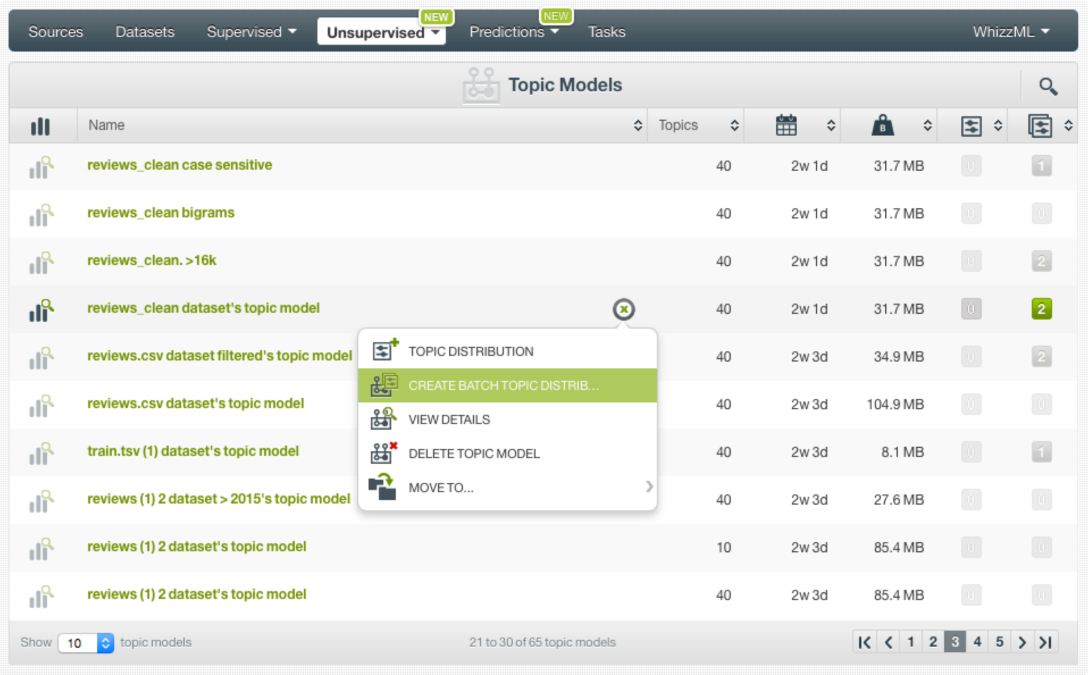


Figure 6.12: Batch topic distribution from 1-click action menu

Alternatively, click CREATE BATCH TOPIC DISTRIBUTION in the **pop up menu** from the topic model list view as shown in Figure 6.13.



Name	Topics				
reviews_clean case sensitive	40	2w 1d	31.7 MB	0	1
reviews_clean bigrams	40	2w 1d	31.7 MB	0	0
reviews_clean. >16k	40	2w 1d	31.7 MB	0	2
reviews_clean dataset's topic model	40	2w 1d	31.7 MB	0	2
reviews.csv dataset filtered's topic model	40	2w 3d	34.9 MB	0	2
reviews.csv dataset's topic model	40	2w 3d	104.9 MB	0	0
train.tsv (1) dataset's topic model	40	2w 3d	8.1 MB	0	1
reviews (1) 2 dataset > 2015's topic model	40	2w 3d	27.6 MB	0	0
reviews (1) 2 dataset's topic model	10	2w 3d	85.4 MB	0	0
reviews (1) 2 dataset's topic model	40	2w 3d	85.4 MB	0	0

Show 10 topic models 21 to 30 of 65 topic models 1 2 3 4 5

Figure 6.13: Batch topic distribution from pop up menu

2. **Select the dataset** containing all the instances you want to predict. (See Figure 6.14.) The instances should contain the input text for the fields used by the topic model. From this view you can also select another topic model using the selector.

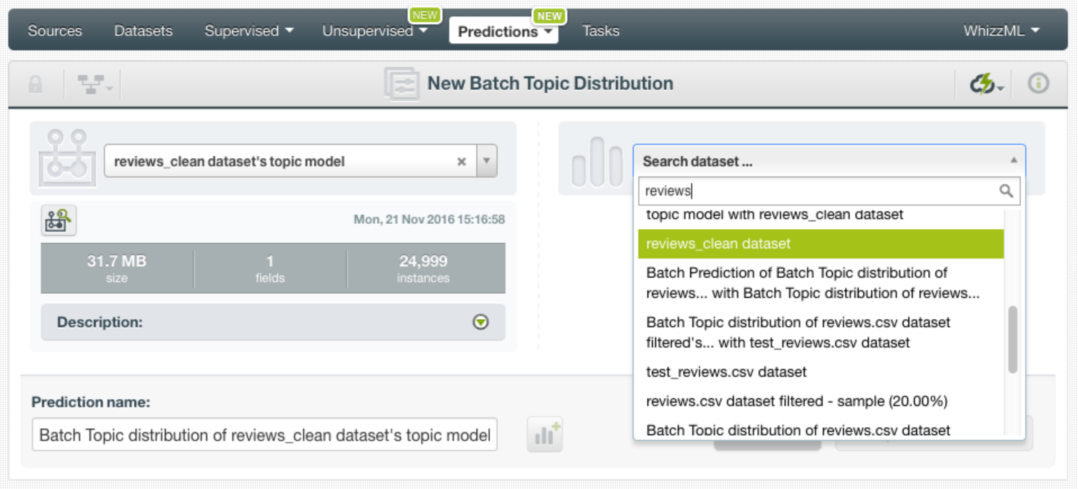
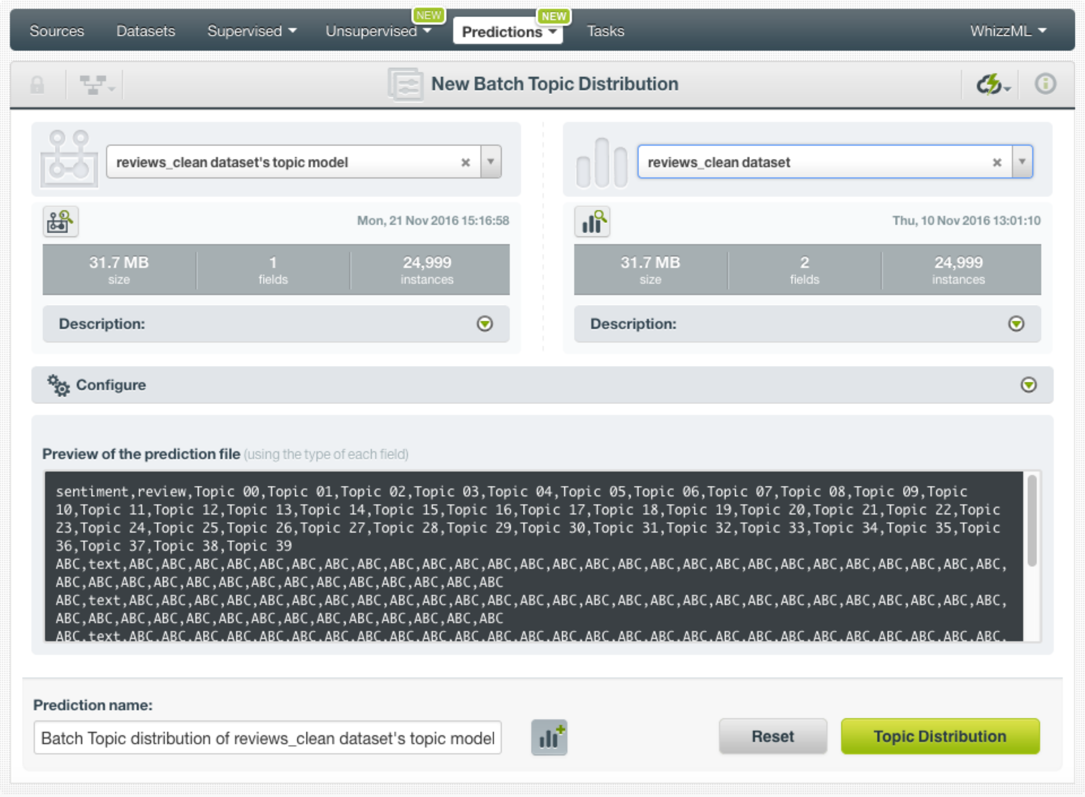

Sources
Datasets
Supervised
Unsupervised
NEW
Predictions
NEW
Tasks
WhizzML
New Batch Topic Distribution
reviews_clean dataset's topic model
Mon, 21 Nov 2016 15:16:58
31.7 MB
size
1
fields
24,999
instances
Description:
Search dataset ...
reviews
topic model with reviews_clean dataset
reviews_clean dataset
Batch Prediction of Batch Topic distribution of reviews... with Batch Topic distribution of reviews...
Batch Topic distribution of reviews.csv dataset filtered's... with test_reviews.csv dataset
test_reviews.csv dataset
reviews.csv dataset filtered - sample (20.00%)
Batch Topic distribution of reviews.csv dataset
Prediction name:
Batch Topic distribution of reviews_clean dataset's topic model

Figure 6.14: Select dataset for batch topic distributions

Note: BigML batch topic distribution can handle missing data in your dataset.

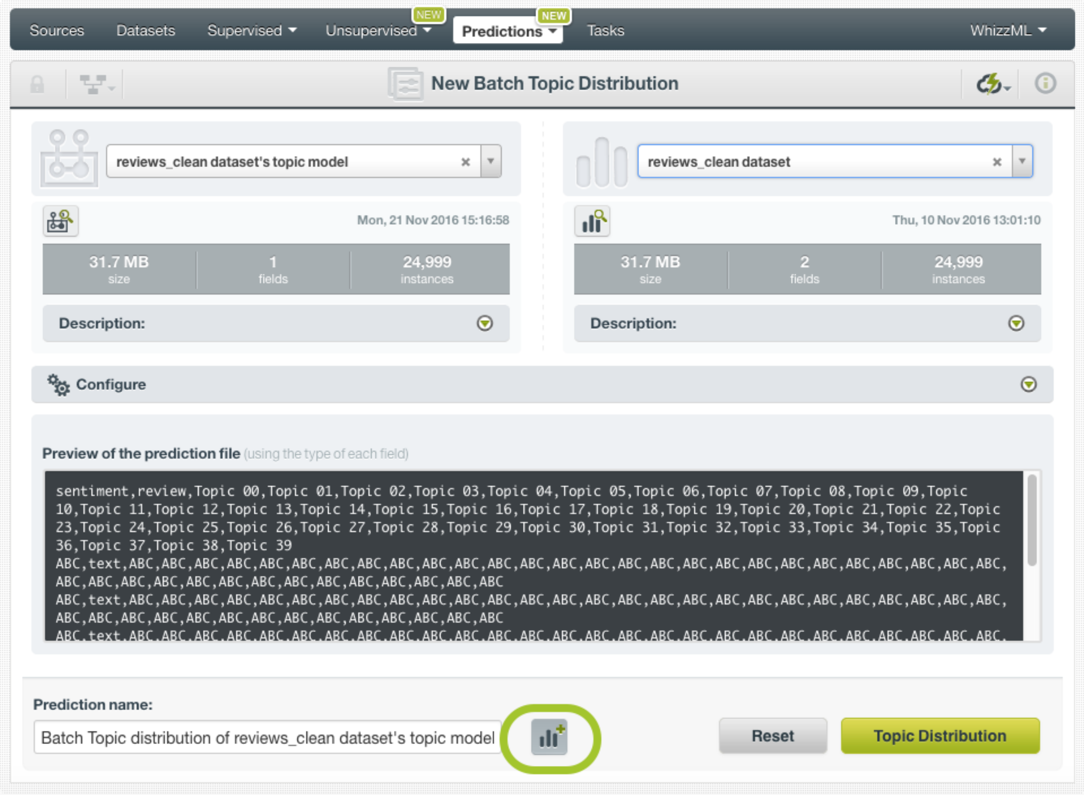
3. After you select the topic model and the dataset, the batch topic distribution **configuration options** will appear along with a **preview of the prediction file**. (See Figure 6.15.) The default format is a CSV file including all your dataset fields and adding extra columns, one per topic, containing the topic probabilities. You can configure this file using the output settings explained in Section 6.3.


Sources
Datasets
Supervised
Unsupervised
NEW
Predictions
NEW
Tasks
WhizzML
New Batch Topic Distribution
reviews_clean dataset's topic model
Mon, 21 Nov 2016 15:16:58
31.7 MB
size
1
fields
24,999
instances
Description:
reviews_clean dataset
Thu, 10 Nov 2016 13:01:10
31.7 MB
size
2
fields
24,999
instances
Description:
Configure
Preview of the prediction file (using the type of each field)
Prediction name:
Batch Topic distribution of reviews_clean dataset's topic model
Reset
Topic Distribution

```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
```

Figure 6.15: Configuration options displayed and output preview

4. By default, BigML generates an **output dataset** with your topic distributions that you can later find in the datasets section of your BigML Dashboard. This dataset can be very helpful if you wish to use the topics as input fields for another model (classification and regression models, clustering analysis, anomaly detection, etc.). This option is active by default, but you can deactivate it by clicking in the icon shown in Figure 6.16.



New Batch Topic Distribution

reviews_clean dataset's topic model

Mon, 21 Nov 2016 15:16:58

31.7 MB size | 1 fields | 24,999 instances

Description:

reviews_clean dataset

Thu, 10 Nov 2016 13:01:10

31.7 MB size | 2 fields | 24,999 instances

Description:

Configure

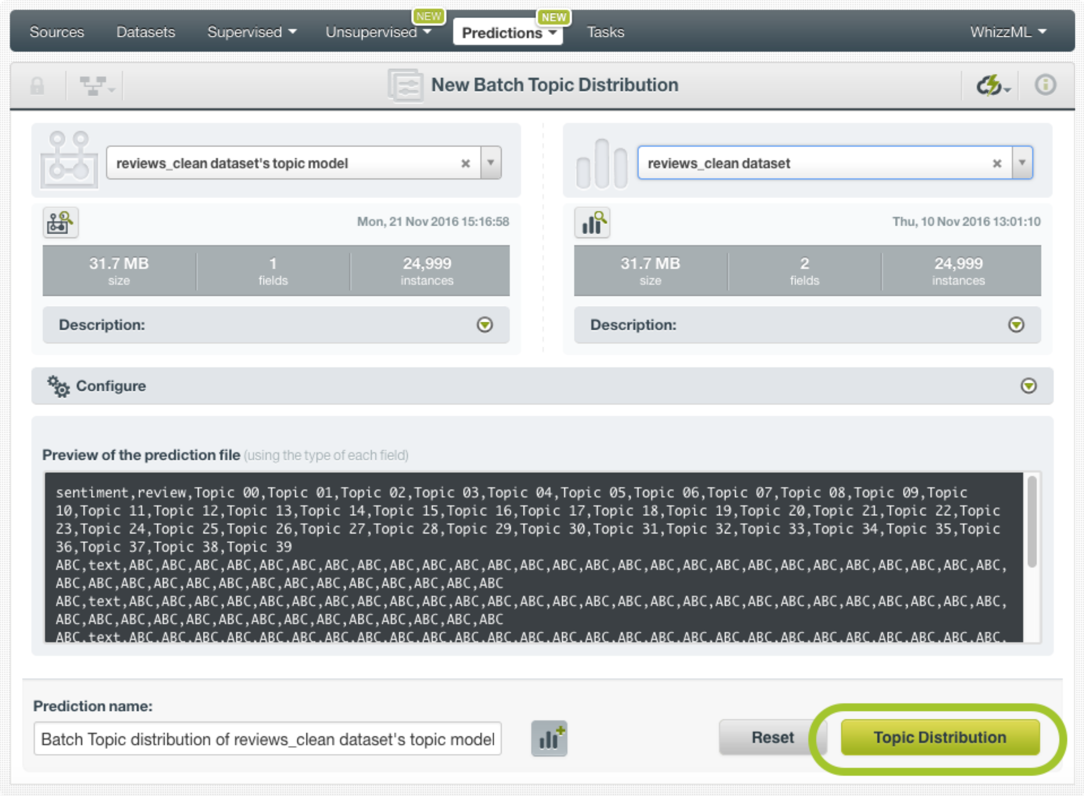
Preview of the prediction file (using the type of each field)

```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
```

Prediction name: Batch Topic distribution of reviews_clean dataset's topic model

Figure 6.16: Create dataset from batch prediction

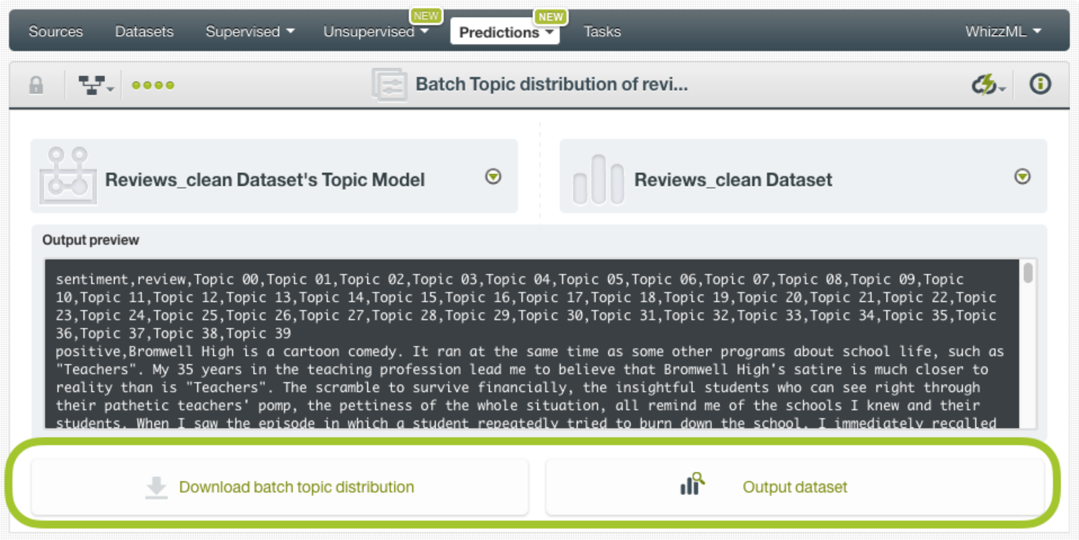
5. Finally, click on the Topic Distribution button to generate your batch topic distribution. (See Figure 6.17.)


Sources
Datasets
Supervised
NEW
Unsupervised
NEW
Predictions
Tasks
WhizzML
New Batch Topic Distribution
reviews_clean dataset's topic model
reviews_clean dataset
Mon, 21 Nov 2016 15:16:58
Thu, 10 Nov 2016 13:01:10
31.7 MB
size
1
fields
24,999
instances
31.7 MB
size
2
fields
24,999
instances
Description:
Description:
Configure
Preview of the prediction file (using the type of each field)
Prediction name:
Batch Topic distribution of reviews_clean dataset's topic model
Reset
Topic Distribution

```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC
ABC,text,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,ABC,
```

Figure 6.17: Click Topic Distribution

6. When the batch topic distribution is created, you will be able to **download the CSV file** containing all your dataset instances along with the topic probabilities. If you did not disable the option to create a new dataset, you will also be able to access the **output dataset** from the batch topic distribution view. (See Figure 6.18.)



```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
positive,Bromwell High is a cartoon comedy. It ran at the same time as some other programs about school life, such as
"Teachers". My 35 years in the teaching profession lead me to believe that Bromwell High's satire is much closer to
reality than is "Teachers". The scramble to survive financially, the insightful students who can see right through
their pathetic teachers' pomp, the pettiness of the whole situation, all remind me of the schools I knew and their
students. When I saw the episode in which a student repeatedly tried to burn down the school, I immediately recalled
```

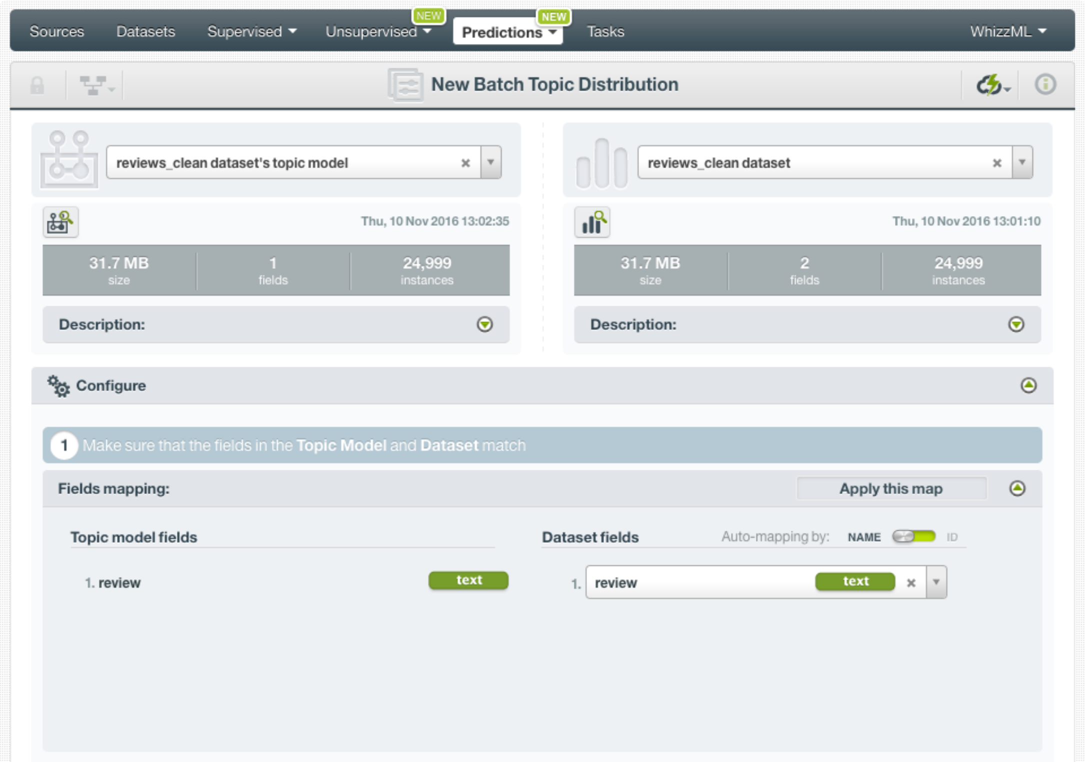
Figure 6.18: Download batch topic distribution and access output dataset

6.3 Configuring Batch Topic Distributions

BigML provides several options to configure your topic distributions, such as defining the automatic **field mapping** performed by BigML and the **output file settings**. See the following sections for an explanation of both options.

6.3.1 Field Mapping

You can specify which fields in the topic model match with which fields in the dataset containing the instances you want to predict. BigML automatically matches fields by **name**, but you can set an automatic match by **field ID** by clicking in the green switcher shown in Figure 6.19. You can also **manually** search for fields or remove them if you do not want to consider them.



reviews_clean dataset's topic model

Thu, 10 Nov 2016 13:02:35

31.7 MB size | 1 fields | 24,999 instances

Description:

reviews_clean dataset

Thu, 10 Nov 2016 13:01:10

31.7 MB size | 2 fields | 24,999 instances

Description:

1 Make sure that the fields in the **Topic Model** and **Dataset** match

1. review text

1. review text

Figure 6.19: Field mapping for batch topic distributions

Note: the field mapping from the BigML Dashboard has a limit of 200 fields, for batch topic distributions with higher number of fields you can use the BigML API[1].

6.3.2 Output Settings

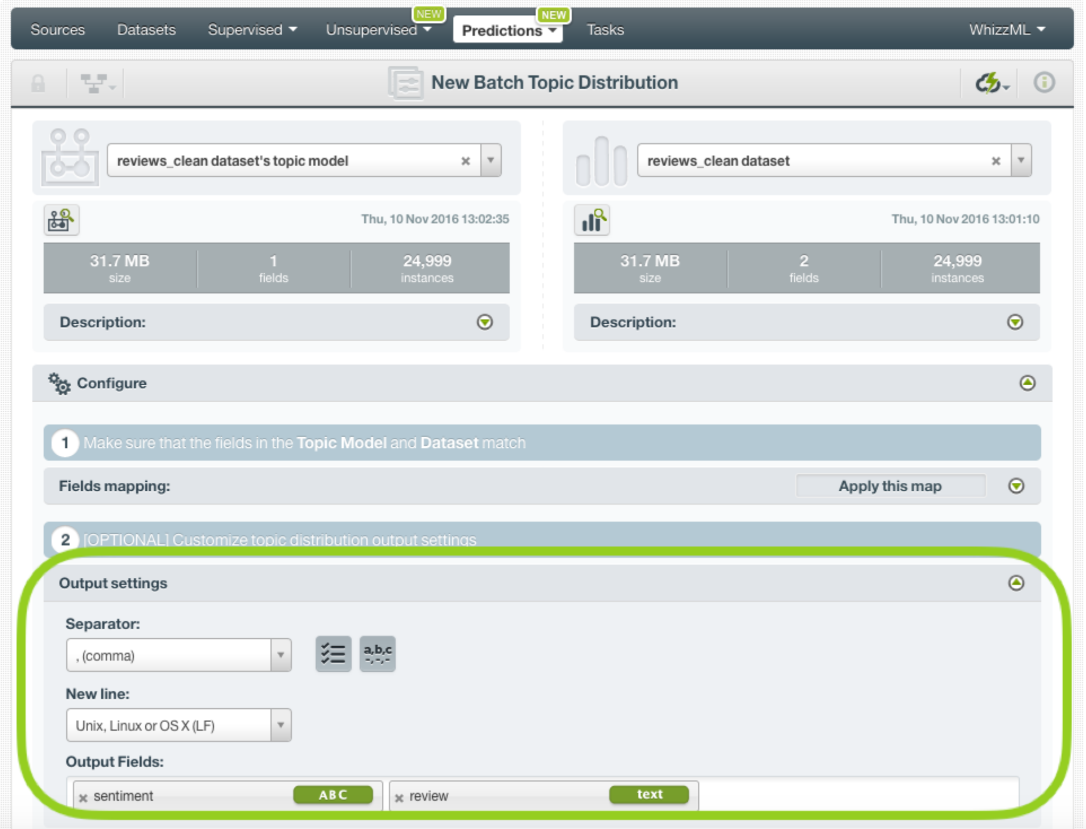
Batch topic distributions return a CSV file containing all the topic probabilities per instance by default. You can tune the following settings to customize your output file:

- **Separator:** this option allows you to choose the best separator for your output file columns. The default separator is comma. You can also select semicolon, tab or space.
- **New line**: this option allows you to set the new line character to use as the line break in the generated csv file: “LF”, “CRLF”.
- **Output fields:** you have an option to include or exclude all your dataset fields from your output file. You can also select the fields you want to include or exclude one by one from the preview shown in Figure 6.20.

  Note: a maximum of 100 fields are displayed in the preview, but all your dataset fields are included in the output file by default unless you exclude them.

[1] https://bigml.com/api/batchtopicdistributions

- **Headers:** this option includes or excludes a first row in the output file (and in the output dataset) with the names of each column. By default, BigML includes the headers.
- **New line**: The new line character that you want to get as line break in the generated CSV file, e.g. “LF” or “CRLF”. It is “LF” by default.



reviews_clean dataset's topic model

Thu, 10 Nov 2016 13:02:35

31.7 MB size | 1 fields | 24,999 instances

Description:

reviews_clean dataset

Thu, 10 Nov 2016 13:01:10

31.7 MB size | 2 fields | 24,999 instances

Description:

Configure

1 Make sure that the fields in the **Topic Model** and **Dataset** match

Fields mapping: Apply this map

2 [OPTIONAL] Customize topic distribution output settings

Output settings

Separator: , (comma)

New line: Unix, Linux or OS X (LF)

Output Fields: sentiment ABC | review text

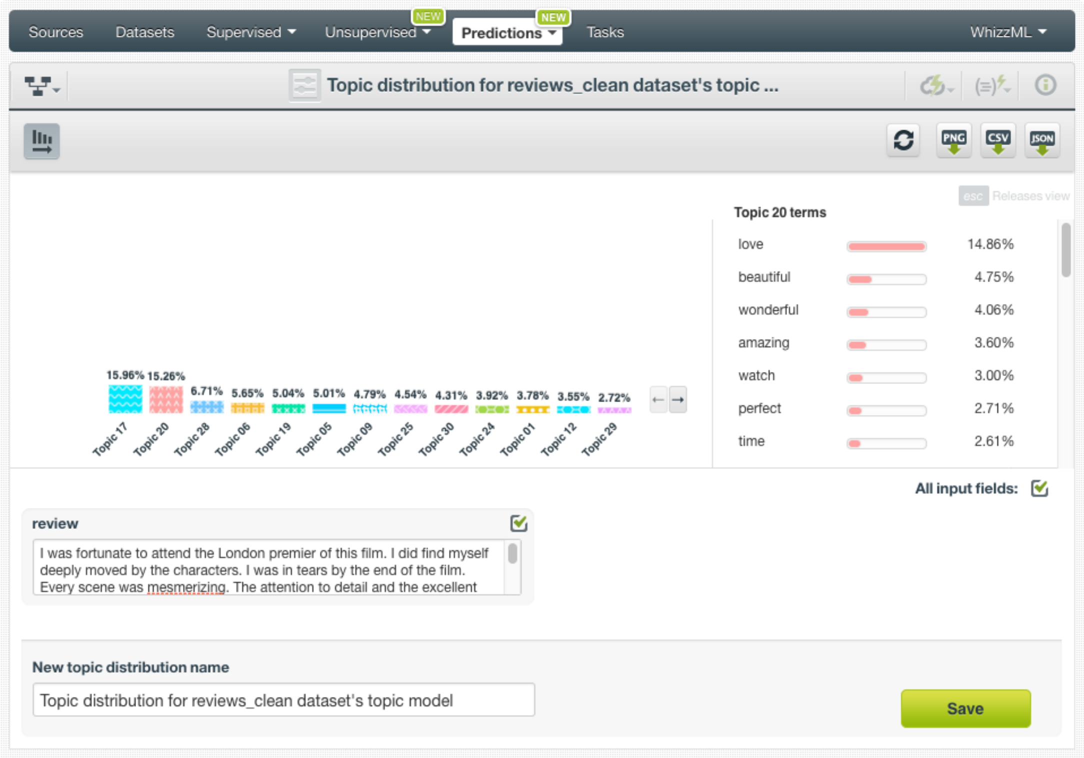
Figure 6.20: Output file settings for batch topic dsitributions

6.4 Visualizing Topic Distributions

The visualization of topic distributions is different for topic distributions for one instance and for batch topic dsitributions. The following sections explain both of them.

6.4.1 Topic Distributions

You can find a **histogram** with **topic probabilities** at the top of the topic distribution form when you insert the text for your input fields. Topics are sorted by probability from left to right, you can sort them by name simply clicking in the sorting option on the top menu. The higher the probability, the more important the topic for the given text. By mousing over a topic bar, you will be able to see the topic **top terms** and their probabilities to the left of the histogram. (See Figure 6.21.)



Topic 20 terms

love	14.86%
beautiful	4.75%
wonderful	4.06%
amazing	3.60%
watch	3.00%
perfect	2.71%
time	2.61%

All input fields: ☑

review ☑

I was fortunate to attend the London premier of this film. I did find myself deeply moved by the characters. I was in tears by the end of the film. Every scene was mesmerizing. The attention to detail and the excellent

New topic distribution name

Topic distribution for reviews_clean dataset's topic model

Save

Figure 6.21: Topic distributions view

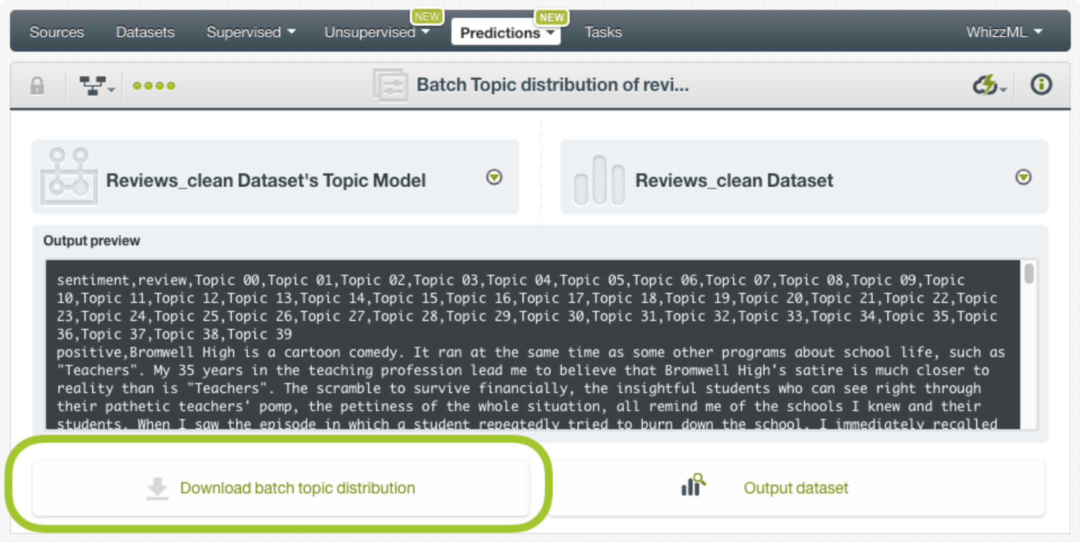
If you click the Save button your topic distributions will be saved and listed in the topic distribution list view. (See Figure 6.1.)

6.4.2 Batch Topic Distributions

For batch topic distributions, you will always get an **output file** and, optionally, an **output dataset**.

Output File

Access the CSV file containing all topic probabilities from the batch topic distribution view, as shown in Figure 6.22.


Sources
Datasets
Supervised
NEW
Unsupervised
NEW
Predictions
Tasks
WhizzML
Batch Topic distribution of revi...
Reviews_clean Dataset's Topic Model
Reviews_clean Dataset
Download batch topic distribution
Output dataset

Output preview

```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
positive,Bromwell High is a cartoon comedy. It ran at the same time as some other programs about school life, such as
"Teachers". My 35 years in the teaching profession lead me to believe that Bromwell High's satire is much closer to
reality than is "Teachers". The scramble to survive financially, the insightful students who can see right through
their pathetic teachers' pomp, the pettiness of the whole situation, all remind me of the schools I knew and their
students. When I saw the episode in which a student repeatedly tried to burn down the school, I immediately recalled
```

Figure 6.22: Download output file

By default, it will be a CSV file containing all the dataset fields and the set of topic probabilities for each of the instances. You can customize the output file settings as explained in Subsection 6.3.2.

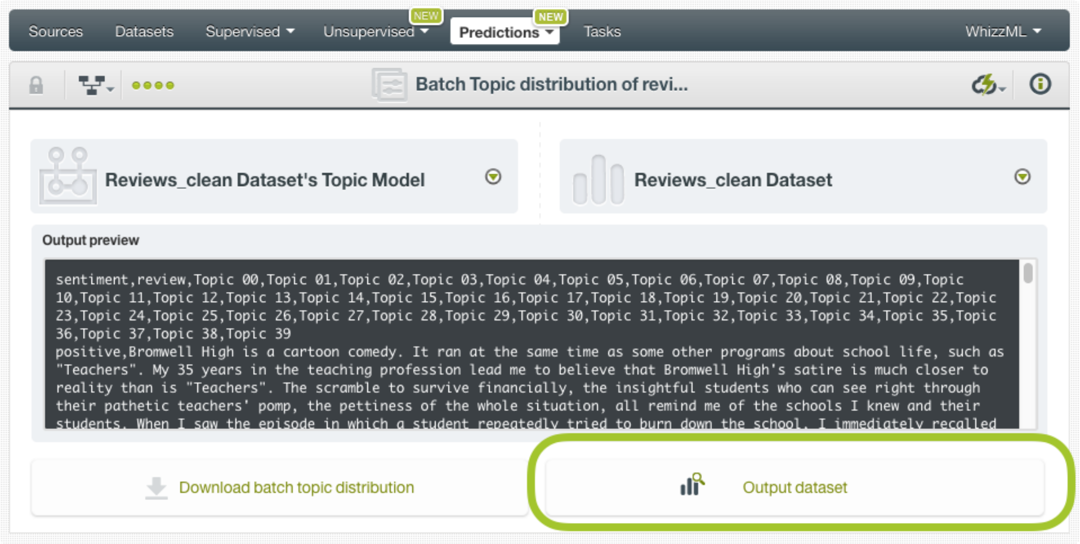
See an output CSV file example in Figure 6.23 where the first column contains the tweets content and the last columns contain the topic distributions for each instance.

```
Tweet, Topic 00, Topic 01, Topic 02, Topic 03, Topic 04, Topic 05, Topic 06
RT @mention Google to Launch Major New Social Network Called Circles, Possibly
Today @mention #SocialMedia, 0.00218, 0.00257, 0.00761, 0.3121, 0.68633,
2.4E-4, 0.00023
Hey @mention why not roll a tractor trailer or 2, full of iPads into #SXSW, I
bet you would sell the entire inventory out! #justsaying, 0.07276, 2.4E-4,
0.21671, 0.3354, 0.41497, 0.02368, 0.2763
I think I might go all weekend without seeing the same iPad case twice...
#sxsw, 0.5216, 7.9E-4, 0.40042, 7.9E-4, 7.9E-4, 0.07015, 7.9E-4
I am suffering from iPad envy., 0.01633, 6.0E-4, 0.02395, 0.81918, 6.0E-4,
0.17838, 6.0E-4
```

Figure 6.23: An example of a batch topic distribution CSV file

Output Dataset

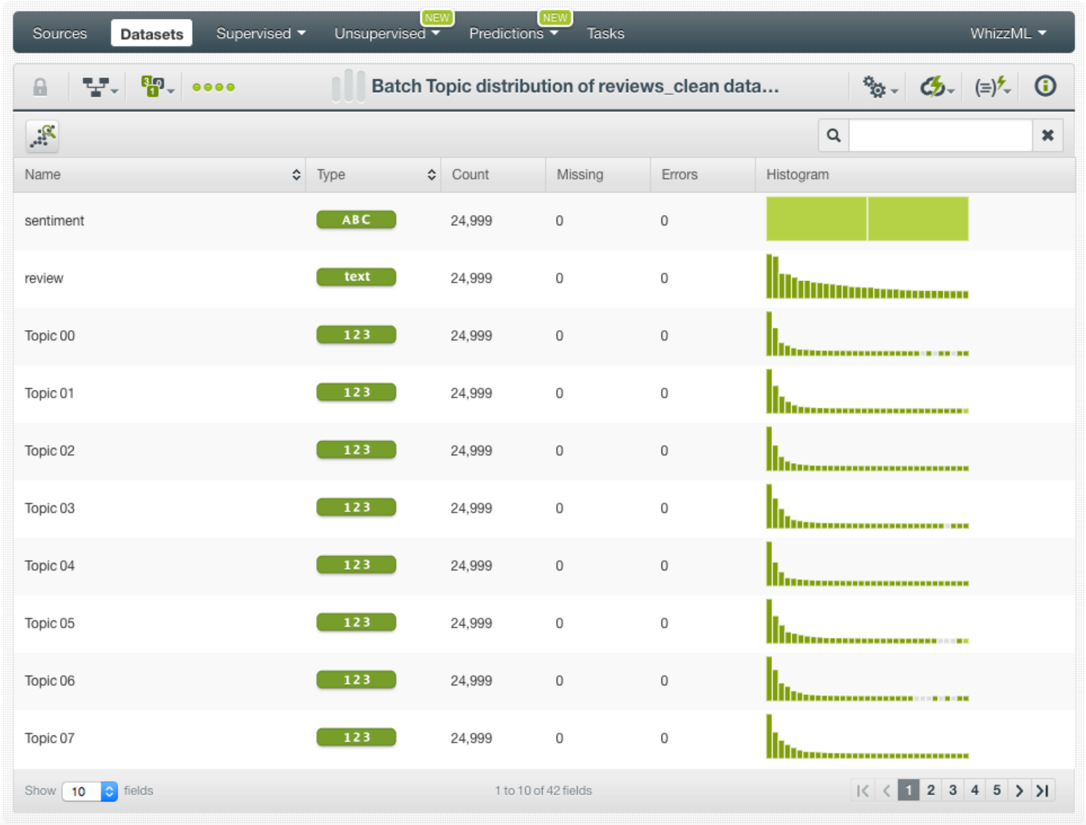
By default BigML automatically creates a dataset out of your batch topic distribution. You can disable this option by configuring your batch topic distribution, as explained in Section 6.3. You can access your output dataset from the batch topic distribution view . (See Figure 6.24.)



```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
positive,Bromwell High is a cartoon comedy. It ran at the same time as some other programs about school life, such as
"Teachers". My 35 years in the teaching profession lead me to believe that Bromwell High's satire is much closer to
reality than is "Teachers". The scramble to survive financially, the insightful students who can see right through
their pathetic teachers' pomp, the pettiness of the whole situation, all remind me of the schools I knew and their
students. When I saw the episode in which a student repeatedly tried to burn down the school, I immediately recalled
```

Figure 6.24: Access batch output dataset

In the output dataset you can find additional **fields**, one per topic, containing the topic probabilities for each of the instances. (See Figure 6.25.)



Batch Topic distribution of reviews_clean data...

Name	Type	Count	Missing	Errors	Histogram
sentiment	ABC	24,999	0	0	
review	text	24,999	0	0	
Topic 00	123	24,999	0	0	
Topic 01	123	24,999	0	0	
Topic 02	123	24,999	0	0	
Topic 03	123	24,999	0	0	
Topic 04	123	24,999	0	0	
Topic 05	123	24,999	0	0	
Topic 06	123	24,999	0	0	
Topic 07	123	24,999	0	0	

Show 10 fields

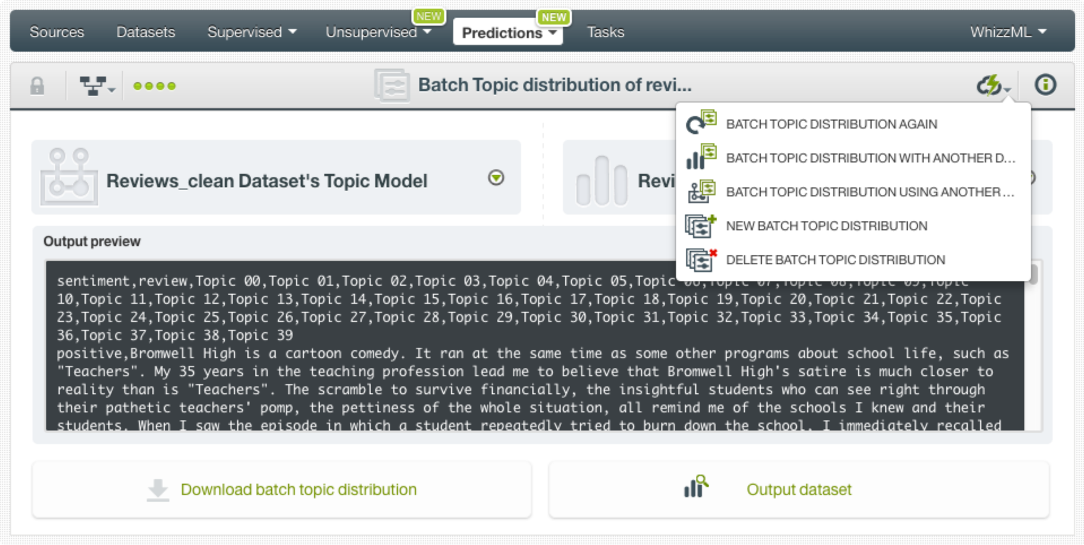
1 to 10 of 42 fields

Figure 6.25: Batch output dataset

Batch Topic Distribution 1-Click Action Menu

From the batch topic distribution view you can perform the following actions shown in Figure 6.26

- BATCH TOPIC DISTRIBUTION AGAIN: redirects you to the batch topic distribution creation view where you will have the same topic model and the same dataset already selected. It is a nice shortcut if you want to create the batch topic distribution again using a different configuration.
- BATCH TOPIC DISTRIBUTION WITH ANOTHER TOPIC MODEL: creates a batch topic distribution using the same topic model and a different dataset.
- BATCH TOPIC DISTRIBUTION USING ANOTHER TOPIC MODEL: creates a batch topic distribution using the same dataset and a different topic model.
- NEW BATCH TOPIC DISTRIBUTION: redirects you to the batch topic distribution creation view where you will be able to select a topic model and a dataset.


Sources
Datasets
Supervised
NEW
Unsupervised
Predictions
NEW
Tasks
WhizzML
Batch Topic distribution of revi...
BATCH TOPIC DISTRIBUTION AGAIN
BATCH TOPIC DISTRIBUTION WITH ANOTHER D...
BATCH TOPIC DISTRIBUTION USING ANOTHER ...
NEW BATCH TOPIC DISTRIBUTION
DELETE BATCH TOPIC DISTRIBUTION
Reviews_clean Dataset's Topic Model
Revi
Output preview
Download batch topic distribution
Output dataset

```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
positive,Bromwell High is a cartoon comedy. It ran at the same time as some other programs about school life, such as
"Teachers". My 35 years in the teaching profession lead me to believe that Bromwell High's satire is much closer to
reality than is "Teachers". The scramble to survive financially, the insightful students who can see right through
their pathetic teachers' pomp, the pettiness of the whole situation, all remind me of the schools I knew and their
students. When I saw the episode in which a student repeatedly tried to burn down the school, I immediately recalled
```

Figure 6.26: Batch topic distributions 1-click action menu

6.5 Consuming Topic Distributions

You can fully use topic distributions via the BigML API and bindings. The following subsections explain both tools.

6.5.1 Using Batch Topic Distributions Via the BigML API

You can perform all the actions explained in this document such as creating, configuring, retrieving, listing, updating, and deleting topic distributions via the BigML API.

The example below shows how to create a batch topic distribution after the BIGML_AUTH environment variable that contains your authentication credentials is properly set:

```
curl "https://bigml.io/batchtopicdistribution?$BIGML_AUTH" \
     -X POST \
     -H 'content-type: application/json' \
     -d '{"topicmodel": "topicmodel/5423625af0a5ea3eea000028",
       "dataset": "dataset/54222a14f0a5eaaab000000c"}'
```

For more information on using topic distributions through the BigML API, please refer to topic distribution REST API documentation[2].

6.5.2 Using Topic Distributions Via the BigML bindings

You can also create, configure, retrieve, list, update, and delete single and topic distributions via **BigML bindings** which are libraries aimed to make it easier to use the BigML API from your language of choice. BigML offers bindings in multiple languages including Python, Node.js Java, Swift and Objective-C. See below an example for creating a batch topic distribution with the Python bindings.

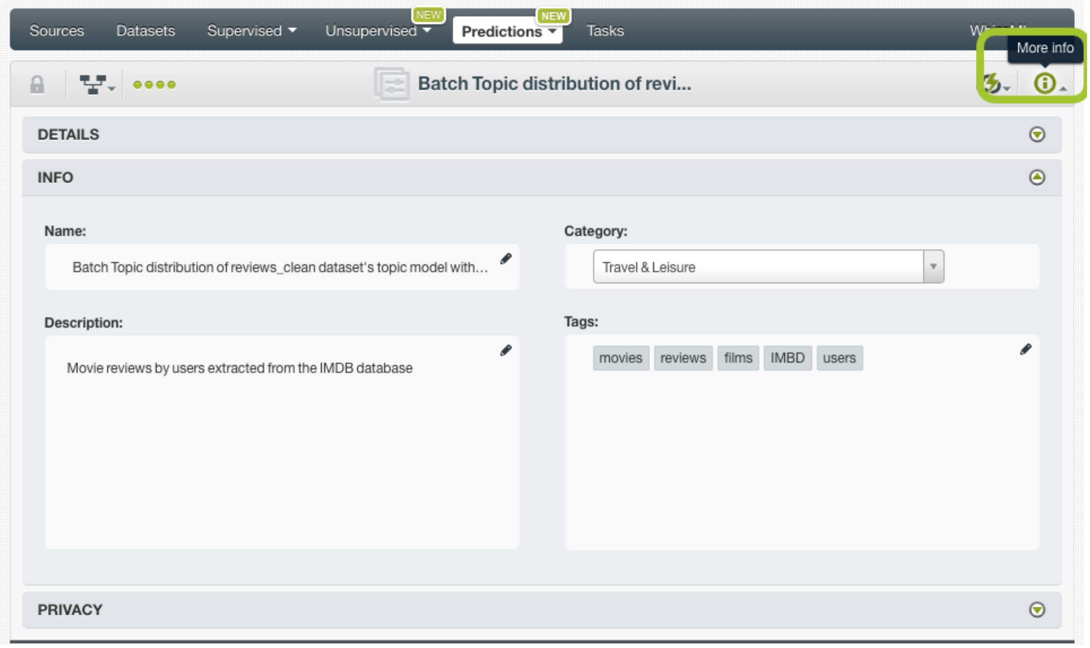
```
from bigml.api import BigML
api = BigML()
batch_topicdistribution = api.create_batch_topic_distribution(
"topicmodel/50650bdf3c19201b64000020",
"dataset": "dataset/54222a14f0a5eaaab000000c")
```

[2] https://bigml.com/api/topicdistributions

For more information on BigML bindings, please refer to the bindings page[3].

6.6 Descriptive Information

Each topic distribution has an associated **name**, **description**, **category** and **tags**. Those options are editable through the MORE INFO menu on the top right of the topic distribution view. (See Figure 6.27.)



Name: Batch Topic distribution of reviews_clean dataset's topic model with...

Category: Travel & Leisure

Description: Movie reviews by users extracted from the IMDB database

Tags: movies reviews films IMBD users

Figure 6.27: Edit topic distributions metadata from More info panel

6.6.1 Topic Distribution Name

If you do not specify a **name** for your topic distributions, BigML assigns a default name depending on the type of topic distribution:

- **Topic distribution:** the name always follow the structure "Topic distribution for <topic model name>".
- **Batch topic distribution:** BigML combines your dataset name and the topic model name: "Batch topic distribution of <topic model name> with <dataset name>".

Topic distribution names are displayed on the list view and also on the top bar of a topic distribution view. Topic distribution names are indexed to be used in searches. You can rename your topic distribution at any time from the MORE INFO menu option.

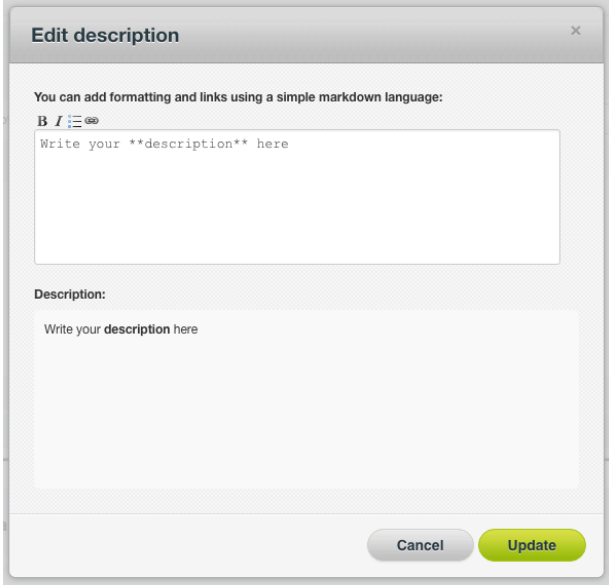
The name of a topic distribution cannot be longer than **256** characters. More than one topic distribution can have the same name even within the same project, but they will always have different identifiers.

6.6.2 Description

Each topic distribution also has a **description** that it is very useful for documenting your Machine Learning projects. Topic distributions take the description from the topic model used to create them.

Descriptions can be written using plain text and also markdown[4]. BigML provides a simple markdown editor that accepts a subset of markdown syntax. (See Figure 6.28.)

[3] https://bigml.com/tools/bindings
[4] https://en.wikipedia.org/wiki/Markdown



You can add formatting and links using a simple markdown language:

```
Write your **description** here
```

Description:

Write your **description** here

Figure 6.28: Markdown editor for topic distributions descriptions

Descriptions cannot be longer than **8192** characters and can use almost any character.

6.6.3 Category

Each topic distribution has associated a **category**. Categories are useful to classify topic distributions according to the domain which your data comes from. This is useful when you use BigML to solve problems across industries or multiple customers. By default, BigML takes the category from the topic distribution used to create it.

A topic distribution category must be one of the categories listed on Table 6.1.

Table 6.1: Categories used to classify topic distributions by BigML

Category
Aerospace and Defense
Automotive, Engineering and Manufacturing
Banking and Finance
Chemical and Pharmaceutical
Consumer and Retail
Demographics and Surveys
Energy, Oil and Gas
Fraud and Crime
Healthcare
Higher Education and Scientific Research
Human Resources and Psychology
Insurance
Law and Order
Media, Marketing and Advertising
Miscellaneous
Physical, Earth and Life Sciences
Professional Services
Public Sector and Nonprofit
Sports and Games
Technology and Communications
Transportation and Logistics
Travel and Leisure
Uncategorized
Utilities

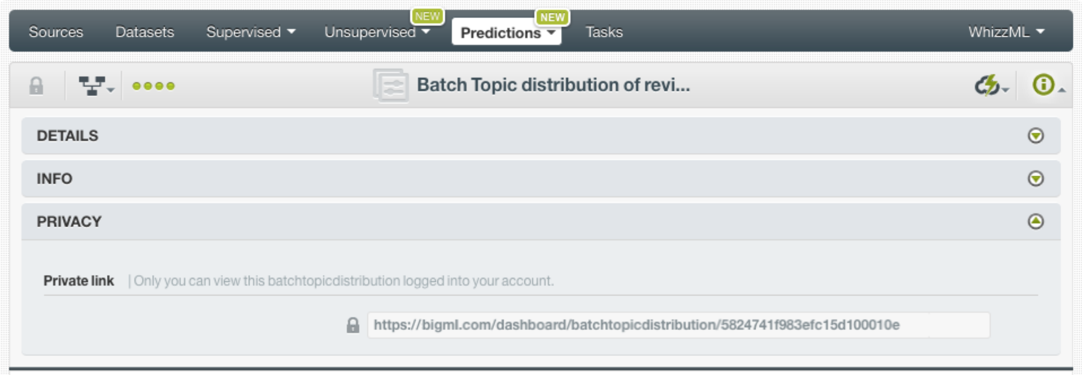
6.6.4 Tags

A topic distribution can also have a number of **tags** associated with it that can help to retrieve it via the BigML API or provide topic distribution with some extra information. A topic distribution inherit the tags from the topic model used to create it.

Each tag is limited to a maximum of 128 characters. Each topic distribution can have up to 32 different tags.

6.7 Topic Distribution Privacy

Privacy options for topic distributions can be defined in the **More Info** menu option. (See Figure 6.29.) There is only one level of privacy for BigML topic distributions: they are private. The link displayed in the **privacy panel** is the private URL of your topic distribution, so only a user logged into your account is able to see it. Neither single nor batch topic distributions can be shared from the BigML Dashboard by sharing a link as you can do with other resources.



Private link | Only you can view this batchtopicdistribution logged into your account.

https://bigml.com/dashboard/batchtopicdistribution/5824741f983efc15d100010e

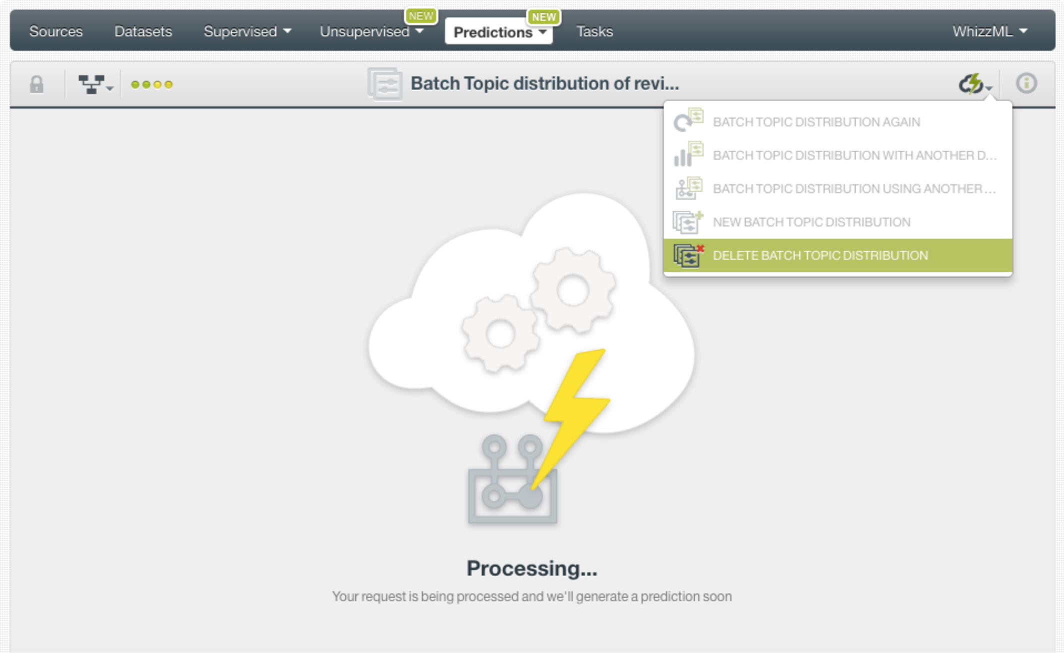
Figure 6.29: Private link of topic distributions

6.8 Moving Topic Distributions

When you create a topic distribution it will be assigned to the same project as the original topic model. You cannot move topic distributions between projects as you do with other resources.

6.9 Stopping Topic Distributions

Batch topic distributions are asynchronous resources so you can stop their creation before the task is finished. You can use the DELETE BATCH TOPIC DISTRIBUTION option from the **1-click action menu**. (See Figure 6.30.)

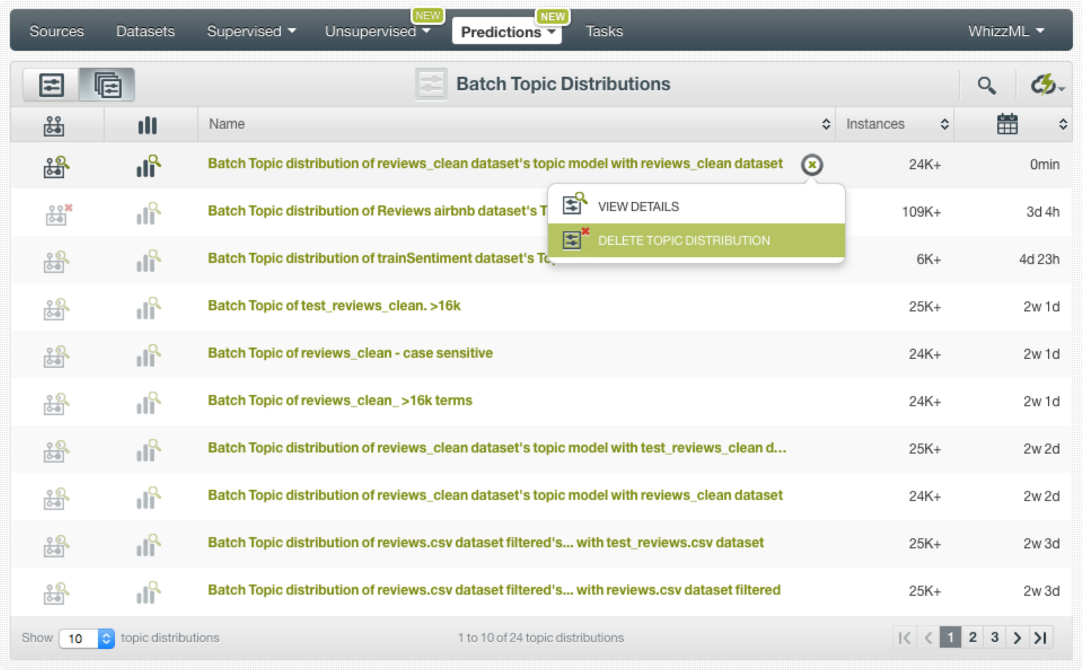

Sources
Datasets
Supervised
Unsupervised
NEW
Predictions
NEW
Tasks
WhizzML
Batch Topic distribution of revi...
BATCH TOPIC DISTRIBUTION AGAIN
BATCH TOPIC DISTRIBUTION WITH ANOTHER D...
BATCH TOPIC DISTRIBUTION USING ANOTHER ...
NEW BATCH TOPIC DISTRIBUTION
DELETE BATCH TOPIC DISTRIBUTION

Processing...

Your request is being processed and we'll generate a prediction soon

Figure 6.30: Stop batch topic distributions from the 1-click action menu

Alternatively, you can use the DELETE BATCH TOPIC DISTRIBUTION from the **topic distribution menu** on the topic distribution list view. (See Figure 6.31.)



Name	Instances	
Batch Topic distribution of reviews_clean dataset's topic model with reviews_clean dataset	24K+	0min
Batch Topic distribution of Reviews airbnb dataset's T	109K+	3d 4h
Batch Topic distribution of trainSentiment dataset's To	6K+	4d 23h
Batch Topic of test_reviews_clean. >16k	25K+	2w 1d
Batch Topic of reviews_clean - case sensitive	24K+	2w 1d
Batch Topic of reviews_clean_ >16k terms	24K+	2w 1d
Batch Topic distribution of reviews_clean dataset's topic model with test_reviews_clean d...	25K+	2w 2d
Batch Topic distribution of reviews_clean dataset's topic model with reviews_clean dataset	24K+	2w 2d
Batch Topic distribution of reviews.csv dataset filtered's... with test_reviews.csv dataset	25K+	2w 3d
Batch Topic distribution of reviews.csv dataset filtered's... with reviews.csv dataset filtered	25K+	2w 3d

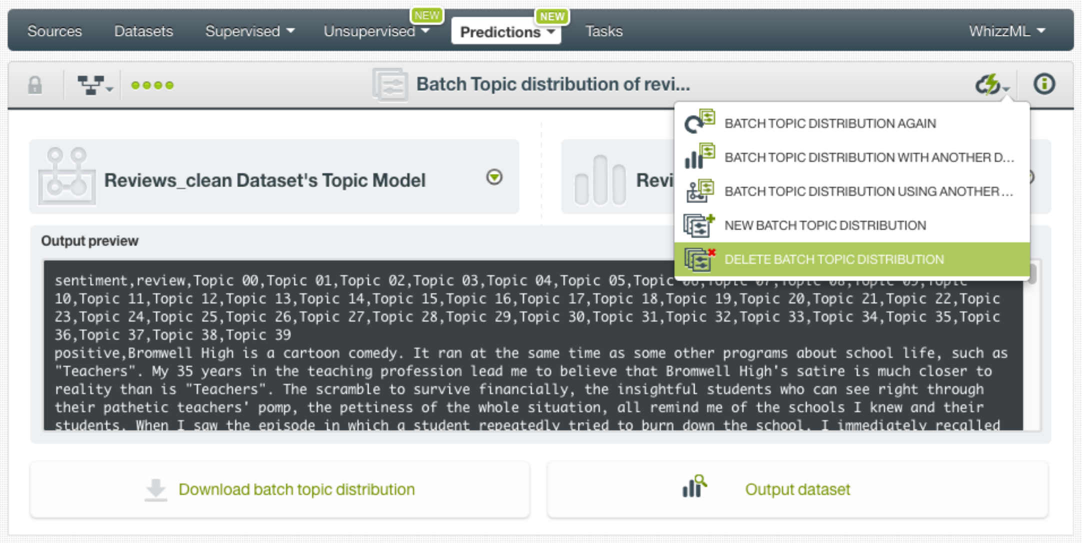
Show 10 topic distributions   1 to 10 of 24 topic distributions   1 2 3

Figure 6.31: Stop batch topic distributions from the pop up menu

Note: if you stop the topic distribution during its creation you will not be able to resume the same task again. If you want to create the same topic distribution you will have to start a new task.

6.10 Deleting Topic Distributions

You can delete your single and batch topic distributions by clicking on the DELETE TOPIC DISTRIBUTION or DELETE BATCH TOPIC DISTRIBUTION option in the **1-click action menu**. (See Figure 6.32).



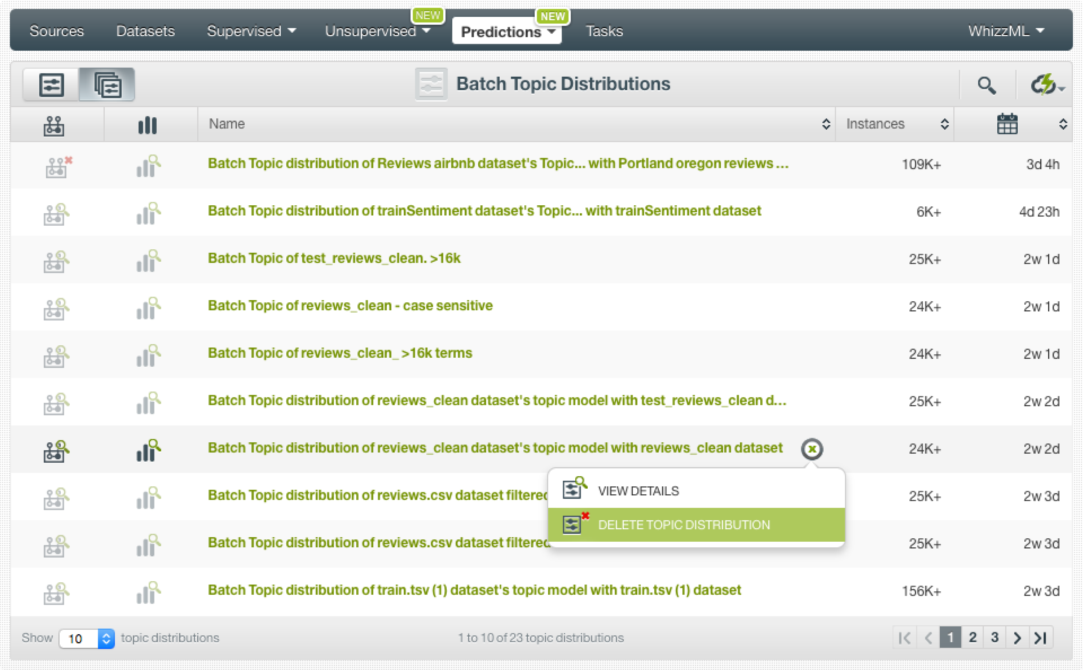
Sources
Datasets
Supervised
NEW
Unsupervised
NEW
Predictions
Tasks
WhizzML
Batch Topic distribution of revi...
BATCH TOPIC DISTRIBUTION AGAIN
BATCH TOPIC DISTRIBUTION WITH ANOTHER D...
BATCH TOPIC DISTRIBUTION USING ANOTHER ...
NEW BATCH TOPIC DISTRIBUTION
DELETE BATCH TOPIC DISTRIBUTION
Reviews_clean Dataset's Topic Model
Revi
Download batch topic distribution
Output dataset

Output preview

```
sentiment,review,Topic 00,Topic 01,Topic 02,Topic 03,Topic 04,Topic 05,Topic 06,Topic 07,Topic 08,Topic 09,Topic
10,Topic 11,Topic 12,Topic 13,Topic 14,Topic 15,Topic 16,Topic 17,Topic 18,Topic 19,Topic 20,Topic 21,Topic 22,Topic
23,Topic 24,Topic 25,Topic 26,Topic 27,Topic 28,Topic 29,Topic 30,Topic 31,Topic 32,Topic 33,Topic 34,Topic 35,Topic
36,Topic 37,Topic 38,Topic 39
positive,Bromwell High is a cartoon comedy. It ran at the same time as some other programs about school life, such as
"Teachers". My 35 years in the teaching profession lead me to believe that Bromwell High's satire is much closer to
reality than is "Teachers". The scramble to survive financially, the insightful students who can see right through
their pathetic teachers' pomp, the pettiness of the whole situation, all remind me of the schools I knew and their
students. When I saw the episode in which a student repeatedly tried to burn down the school, I immediately recalled
```

Figure 6.32: Delete batch topic distributions from the 1-click menu

Alternatively, you can click the DELETE TOPIC DISTRIBUTION in the **pop up menu** from the list view. (See Figure 6.33.)



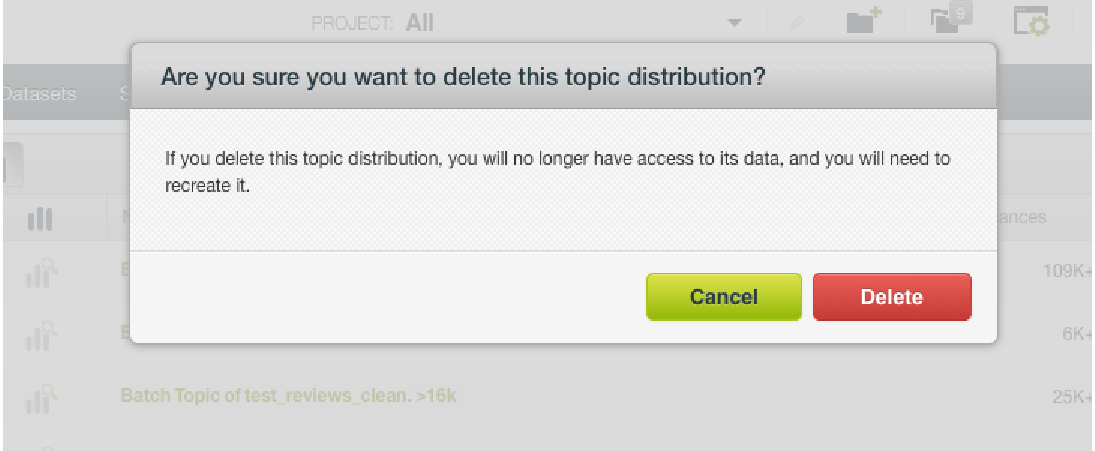
Batch Topic Distributions

Name	Instances	
Batch Topic distribution of Reviews airbnb dataset's Topic... with Portland oregon reviews ...	109K+	3d 4h
Batch Topic distribution of trainSentiment dataset's Topic... with trainSentiment dataset	6K+	4d 23h
Batch Topic of test_reviews_clean. >16k	25K+	2w 1d
Batch Topic of reviews_clean - case sensitive	24K+	2w 1d
Batch Topic of reviews_clean_ >16k terms	24K+	2w 1d
Batch Topic distribution of reviews_clean dataset's topic model with test_reviews_clean d...	25K+	2w 2d
Batch Topic distribution of reviews_clean dataset's topic model with reviews_clean dataset	24K+	2w 2d
Batch Topic distribution of reviews.csv dataset filtered	25K+	2w 3d
Batch Topic distribution of reviews.csv dataset filtered	25K+	2w 3d
Batch Topic distribution of train.tsv (1) dataset's topic model with train.tsv (1) dataset	156K+	2w 3d

Show 10 topic distributions 1 to 10 of 23 topic distributions 1 2 3

Figure 6.33: Delete batch topic distribution from pop up menu

A modal window will be displayed asking you for confirmation. Once a topic distribution is deleted, it is permanently deleted and there is no way you (or even the IT folks at BigML) can retrieve it. (See Figure 6.34.)



Are you sure you want to delete this topic distribution?

If you delete this topic distribution, you will no longer have access to its data, and you will need to recreate it.

Figure 6.34: Delete topic distribution confirmation

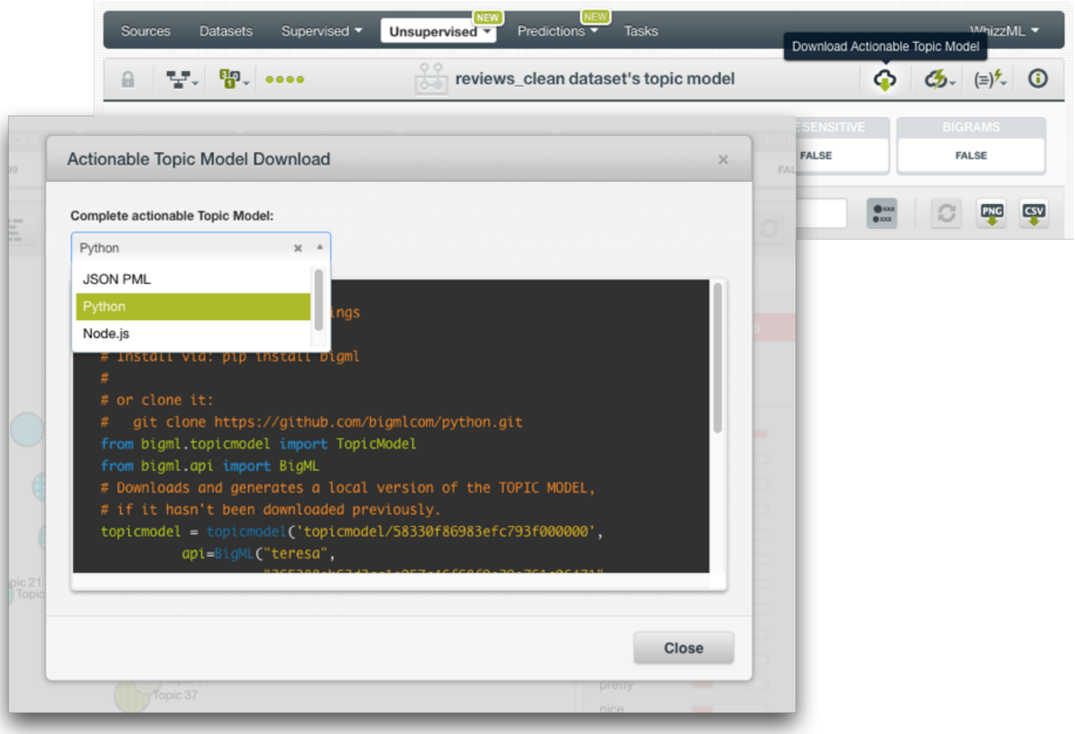
CHAPTER 7

Consuming Topic Models

Similar to other models in BigML, you can **download** topic models and used them locally to make predictions. You can also create and use your topic models programmatically via the **BigML API and bindings**. The following subsections explain those three options.

7.1 Downloading Topic Models

You can download your topic model in a number of languages including your topic model in Python, JSON PML or Node.js. Just click on the DOWNLOAD ACTIONABLE TOPIC MODEL menu option and select your preferred language. (See Figure 7.1.)



```
                                   ings

# install via: pip install bigml
#
# or clone it:
#   git clone https://github.com/bigmlcom/python.git
from bigml.topicmodel import TopicModel
from bigml.api import BigML
# Downloads and generates a local version of the TOPIC MODEL,
# if it hasn't been downloaded previously.
topicmodel = topicmodel('topicmodel/58330f86983efc793f000000',
          api=BigML("teresa",
```

Figure 7.1: Downloading topic model

You can predict the Topic Distribution for your new data locally, free of latency, and at no cost by downloading your topic models. It works the same way as local predictions for models and ensembles.

7.2 Using Topic Models Via the BigML API

Topic models have full citizenship in the BigML API which allows you to programmatically create, configure, retrieve, list, update, delete, and use them to predict topic probabilities for your data.

See in the example below how to create a topic distribution using an existing dataset after you have properly set the BIGML_AUTH environment variable to contain your authentication credentials:

```
curl "https://bigml.io/topicmodel?$BIGML_AUTH" \
     -X POST \
     -H 'content-type: application/json' \
     -d '{"dataset": "dataset/50650bdf3c19201b640542310"}'
```

For more information on using topic models through the BigML API, please refer to topic model REST API documentation[1].

7.3 Using Topic Models Via the BigML Bindings

You can also create and use topic models via **BigML bindings**, which are libraries aimed to make it easier to use the BigML API from your language of choice. BigML offers bindings in multiple languages including Python, Node.js, Java, Swift and Objective-C. You can see below an example to create a topic model with the Python bindings.

```
from bigml.api import BigML
api = BigML()
topicmodel = api.create_topicmodel('dataset/57506c472275c1666b004b10')
```

For more information on BigML bindings, please refer to the bindings page[2].

[1] https://bigml.com/api/topicmodels

[2] https://bigml.com/tools/bindings

CHAPTER 8

Topic Model Limits

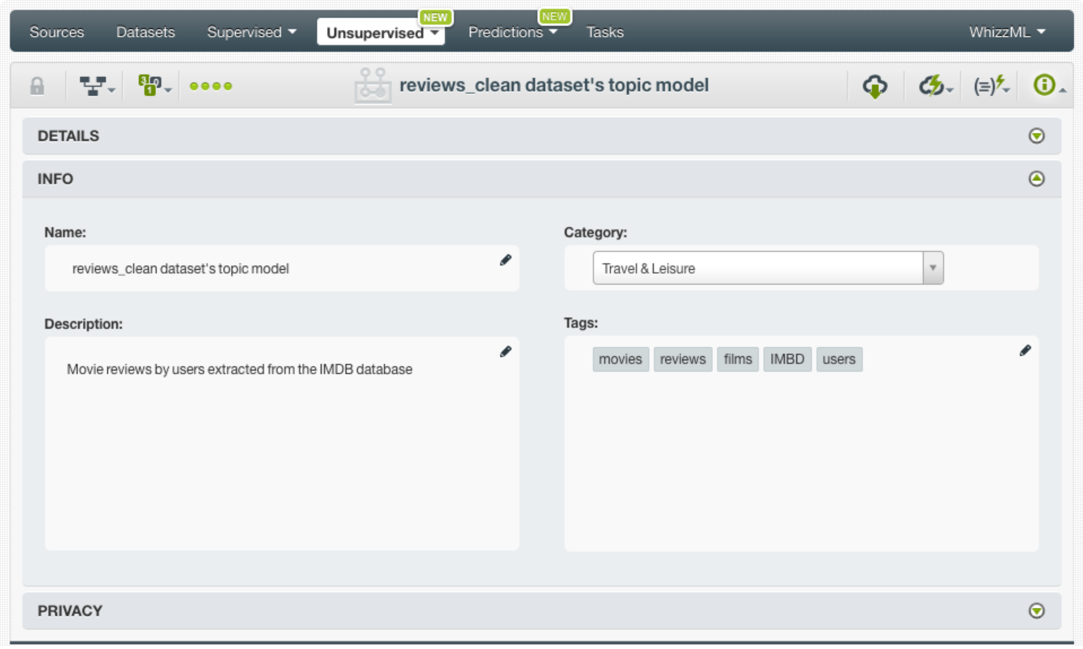
BigML topic models only support **text fields** as inputs. If your dataset contains other field types, those will be ignored by the model. BigML also impose some limits on the configurable parameters to create a topic model. See all limits listed below:

- **Number of topics**: a maximum number of 64 topics is allowed.
- **Number of top terms**: you can choose to display up to 128 terms per topic.
- **Number of terms**: you can select up to 16,384 unique terms to be considered by the model vocabulary.

CHAPTER 9

Topic Model Descriptive Information

Topic models have an associated **name**, **description**, **category**, and **tags**. The following subsections briefly describe each concept. See in Figure 9.1 the options under MORE INFO menu to edit topic models.



Name: reviews_clean dataset's topic model

Category: Travel & Leisure

Description: Movie reviews by users extracted from the IMDB database

Tags: movies reviews films IMBD users

Figure 9.1: Editing topic models

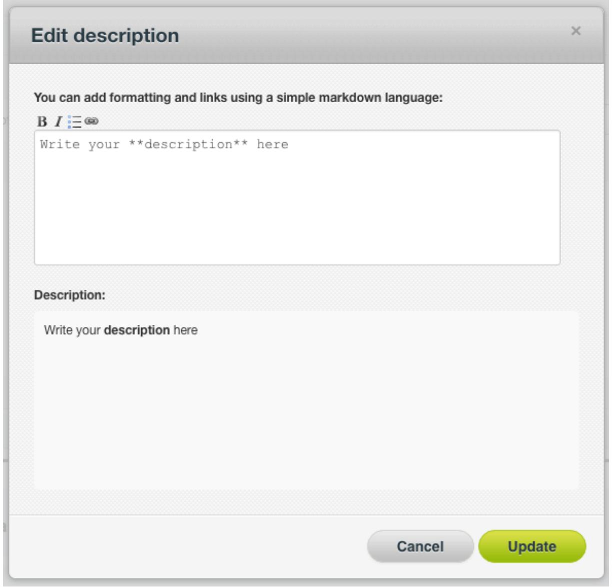
9.1 Topic Model Name

Each topic model has a name that is displayed in the topic model list view and also on the top bar of the topic model view. Topic model names are indexed to be used in searches. When you create a topic model, it gets a default name adding "topic model" at the end of the dataset name: "<dataset name>'s topic model". You can change it using the MORE INFO menu option on the right corner of the topic model view. The name of a topic model cannot be longer than **256** characters. More than one topic model can have the same name even within the same project, but they will always have different identifiers.

9.2 Description

Each topic model also has a **description** that is very useful for documenting your Machine Learning projects. Topic models take the description of the datasets used to create them by default.

Descriptions can be written using plain text and also markdown[1]. BigML provides a simple markdown editor that accepts a subset of markdown syntax. (See Figure 9.2.)



Edit description

You can add formatting and links using a simple markdown language:

```
Write your **description** here
```

Description:

Write your **description** here

Figure 9.2: Markdown editor for topic model descriptions

Descriptions cannot be longer than **8192** characters and can use almost any character.

9.3 Category

A **category**, taken from the dataset used to create it, is associated with each topic model. Categories are useful to classify topic models according to the domain which your data comes from. This is useful when you use BigML to solve problems across industries or multiple customers.

A topic model category must be one of the **24** categories listed in Table 9.1.

[1] https://en.wikipedia.org/wiki/Markdown

Table 9.1: Categories used to classify topic models by BigML

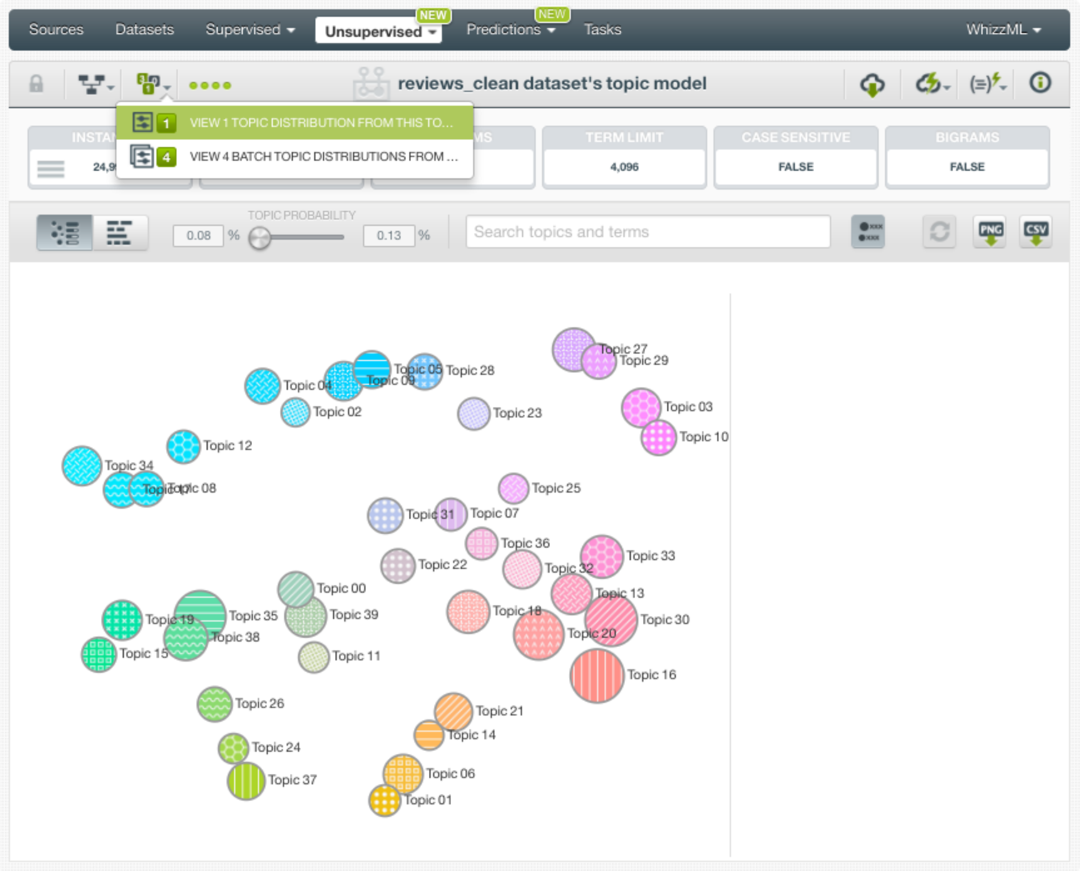
Category
Aerospace and Defense
Automotive, Engineering and Manufacturing
Banking and Finance
Chemical and Pharmaceutical
Consumer and Retail
Demographics and Surveys
Energy, Oil and Gas
Fraud and Crime
Healthcare
Higher Education and Scientific Research
Human Resources and Psychology
Insurance
Law and Order
Media, Marketing and Advertising
Miscellaneous
Physical, Earth and Life Sciences
Professional Services
Public Sector and Nonprofit
Sports and Games
Technology and Communications
Transportation and Logistics
Travel and Leisure
Uncategorized
Utilities

9.4 Tags

A topic model can also have a number of **tags** associated with it that can help to retrieve it via the BigML API or to provide topic models with some extra information. Topic models inherit the tags from the dataset used to create them. Each tag is limited to a maximum of 128 characters. Each topic model can have up to **32** different tags.

9.5 Counters

For each topic model, BigML also stores a number of **counters** to track the number of other resources that have been created using the topic model as a starting point. Display the counters by mousing over the menu option at the top of the topic model view. Click on VIEW # TOPIC DISTRIBUTIONS FROM THIS TOPIC MODEL menu option to quickly access the topic distributions and VIEW # BATCH TOPIC DISTRIBUTIONS FROM THIS TOPIC MODEL to see all batch topic dsitributions. (See Figure 9.3.)



TERM LIMIT 4,096

CASE SENSITIVE FALSE

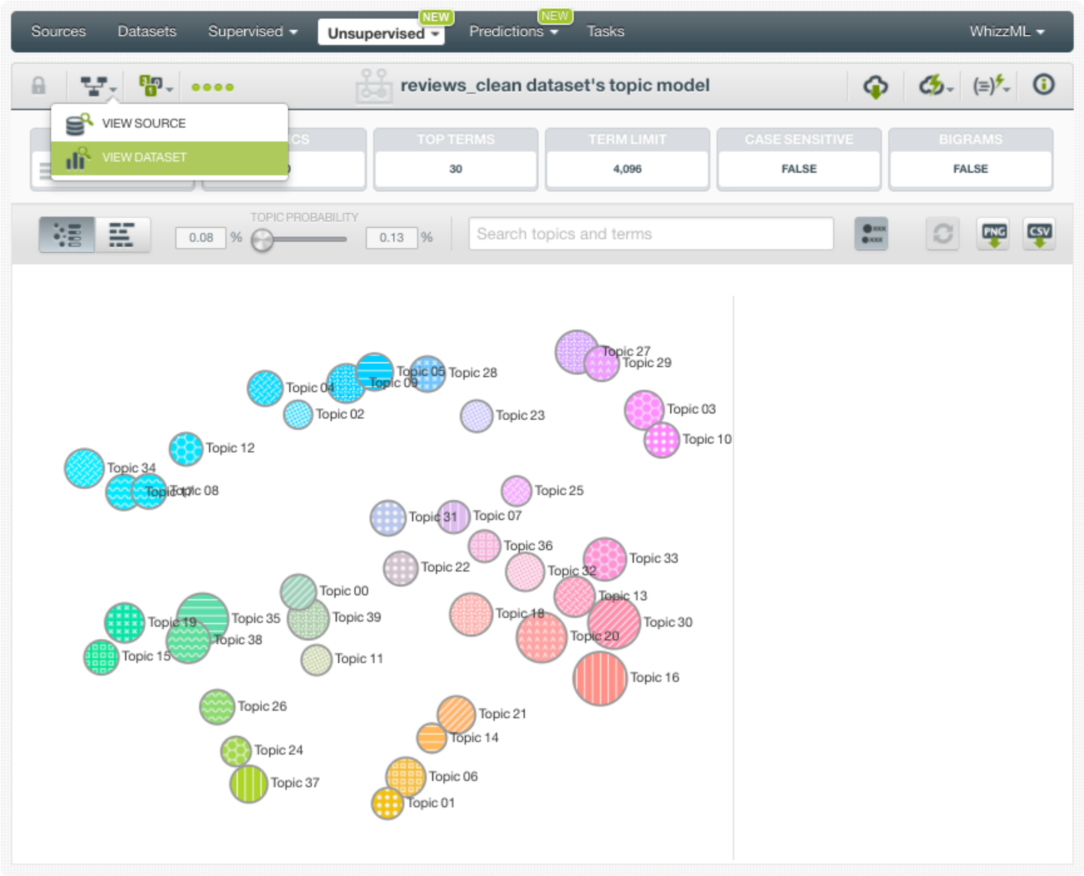
BIGRAMS FALSE

TOPIC PROBABILITY 0.08 % – 0.13 %

Figure 9.3: Counters for topic models

9.6 Original Resources

This menu allows you to keep track of the resources used to create the topic model. You can access the original **source** and the **dataset** by mousing over the icon on the left in the top menu. (See Figure 9.3.) By clicking in the original dataset you will be able to see the **configuration values** used to create your topic model.



TOP TERMS 30

TERM LIMIT 4,096

CASE SENSITIVE FALSE

BIGRAMS FALSE

TOPIC PROBABILITY 0.08 % 0.13 %

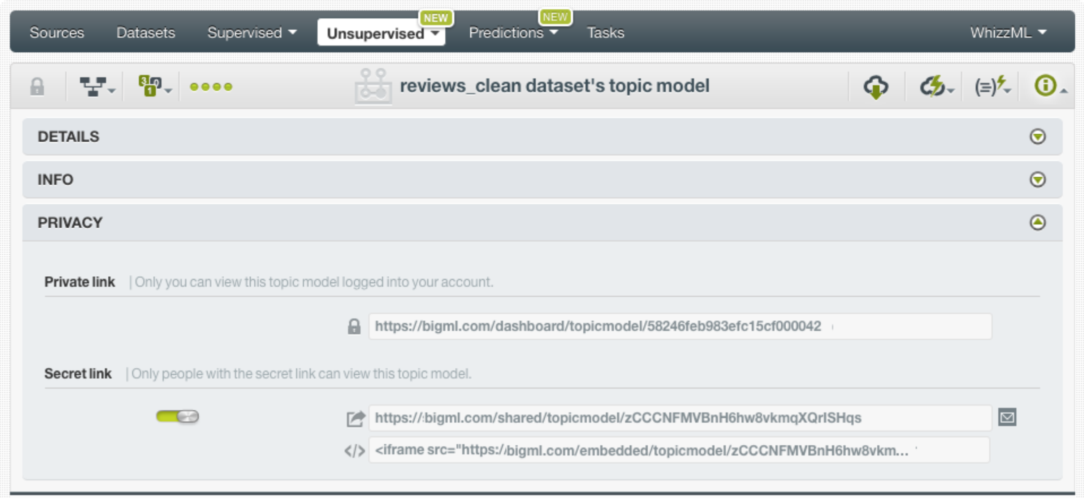
Figure 9.4: Original resources used to create the topic model

CHAPTER 10

Topic Model Privacy

Privacy options for topic models can be defined in the **More Info** menu option. (See Figure 10.1.) There are two levels of privacy for BigML topic models:

- **Private**: only accessible by authorized users (the owner and those who have been granted access by him or her).
- **Shared**: by enabling the **secret link** you will get two different links to share your topic models. The first one is a sharing link that you can copy and send to others so they can visualize and interact with your topic models. The second one is a link to embed your topic models directly on your web page. This is very useful if you want to make local topic distribution predictions at no cost.



Private link | Only you can view this topic model logged into your account.

https://bigml.com/dashboard/topicmodel/58246feb983efc15cf000042

Secret link | Only people with the secret link can view this topic model.

https://bigml.com/shared/topicmodel/zCCCNFMVBnH6hw8vkmqXQrISHqs

<iframe src="https://bigml.com/embedded/topicmodel/zCCCNFMVBnH6hw8vkm...

Figure 10.1: Topic models privacy

CHAPTER 11

Moving Topic Models

When you create a topic model, it will be assigned to the same project as the original dataset. Topic models can only be assigned to a single project. However, you can move topic models between projects. The menu option to do this can be found in two places:

1. Click MOVE TO... within the **1-click action menu** from the topic model view. (See Figure 11.1.)

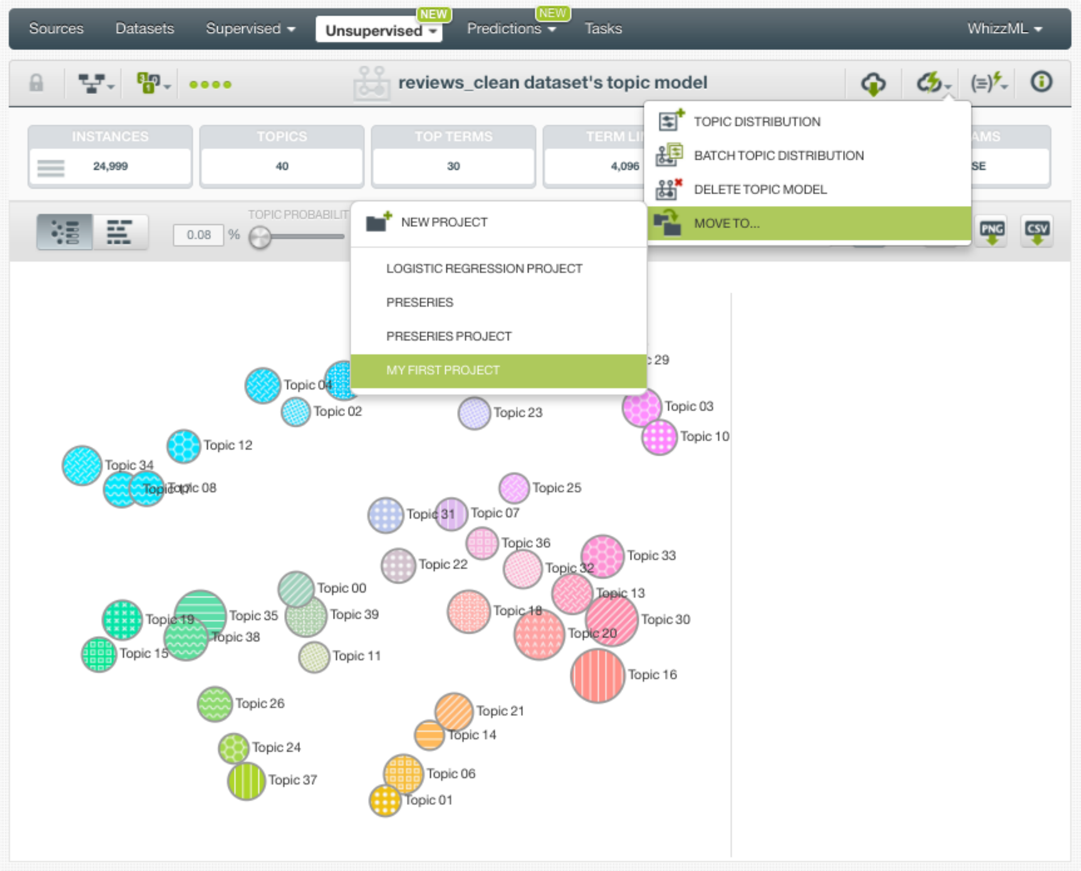
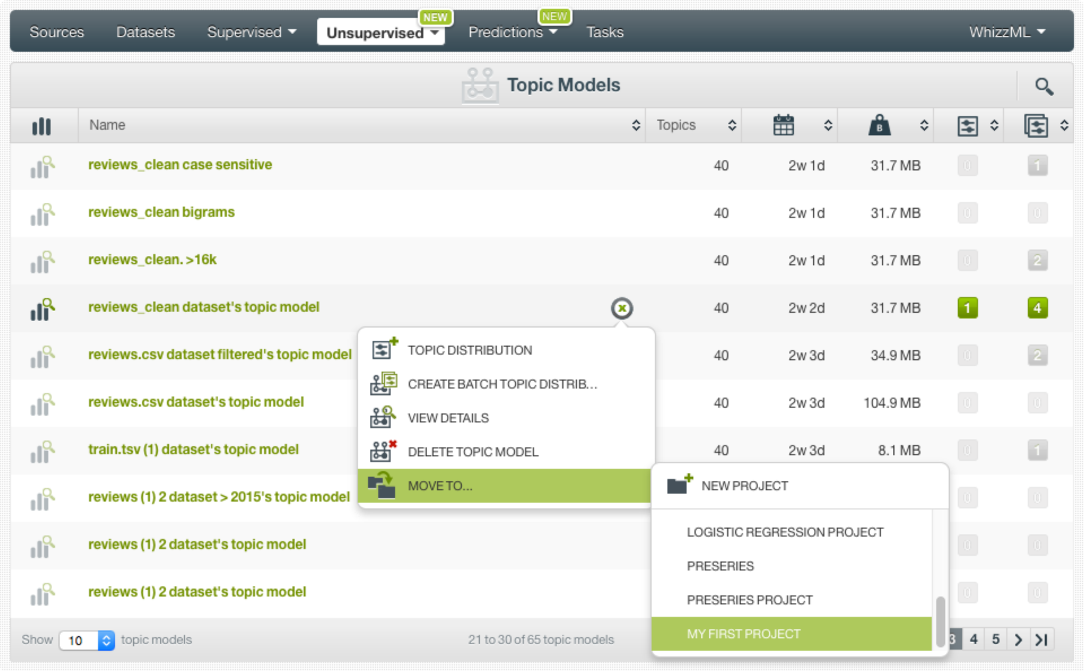


Figure 11.1: Change project from 1-click menu

2. Click MOVE TO... within the **pop up menu** from the topic model list view. (See Figure 11.2)



Name	Topics				
reviews_clean case sensitive	40	2w 1d	31.7 MB	0	1
reviews_clean bigrams	40	2w 1d	31.7 MB	0	0
reviews_clean. >16k	40	2w 1d	31.7 MB	0	2
reviews_clean dataset's topic model	40	2w 2d	31.7 MB	1	4
reviews.csv dataset filtered's topic model	40	2w 3d	34.9 MB	0	2
reviews.csv dataset's topic model	40	2w 3d	104.9 MB	0	0
train.tsv (1) dataset's topic model	40	2w 3d	8.1 MB	0	1
reviews (1) 2 dataset > 2015's topic model					0
reviews (1) 2 dataset's topic model					0
reviews (1) 2 dataset's topic model					0

Figure 11.2: Change project from pop up menu

CHAPTER 12

Stopping Topic Models Creation

You can stop the creation of a topic model before the task is finished by clicking the DELETE TOPIC MODEL option in the **1-click action menu** from the topic model view. (See Figure 12.1.)

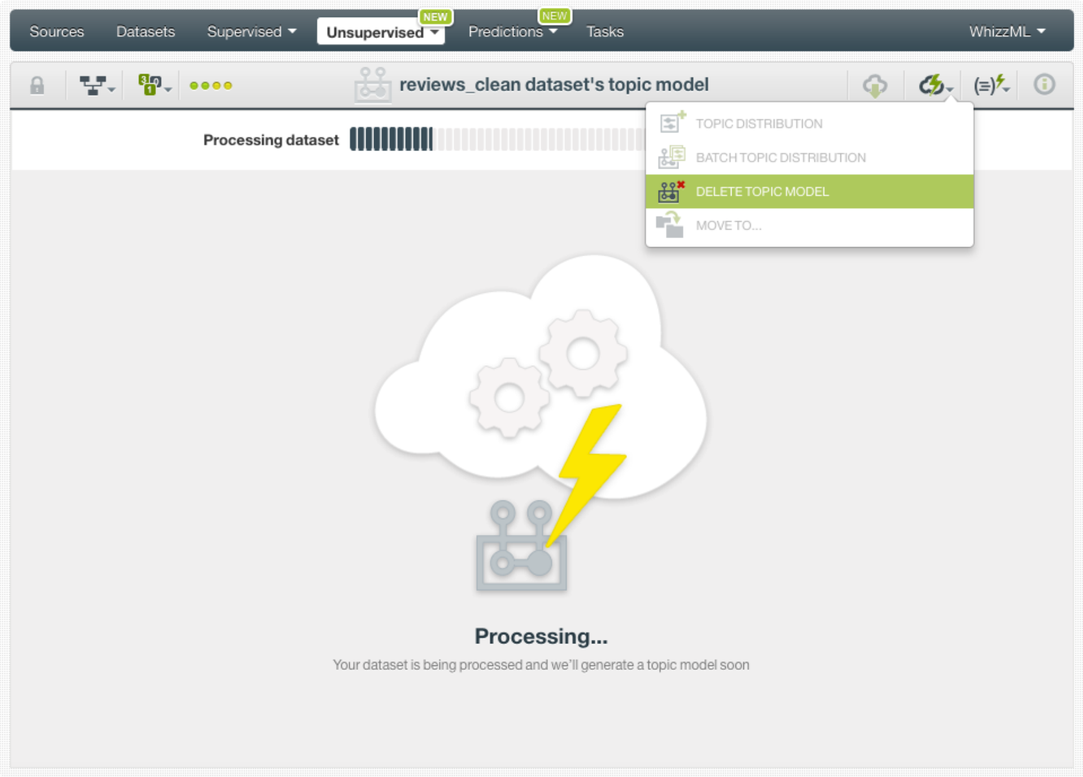
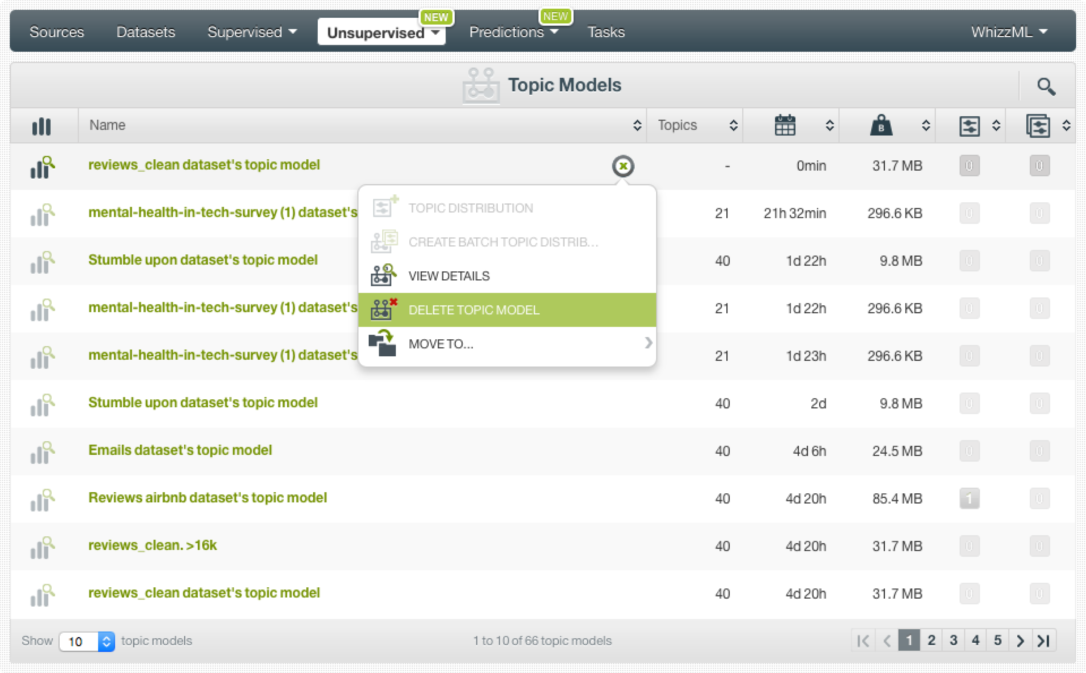


Figure 12.1: Stop topic model creation from 1-click action menu

Alternatively, click the DELETE TOPIC MODEL in the **pop up menu** from the topic model list view. (See Figure 12.2.)



Name	Topics				
reviews_clean dataset's topic model	-	0min	31.7 MB	0	0
mental-health-in-tech-survey (1) dataset's	21	21h 32min	296.6 KB	0	0
Stumble upon dataset's topic model	40	1d 22h	9.8 MB	0	0
mental-health-in-tech-survey (1) dataset's	21	1d 22h	296.6 KB	0	0
mental-health-in-tech-survey (1) dataset's	21	1d 23h	296.6 KB	0	0
Stumble upon dataset's topic model	40	2d	9.8 MB	0	0
Emails dataset's topic model	40	4d 6h	24.5 MB	0	0
Reviews airbnb dataset's topic model	40	4d 20h	85.4 MB	1	0
reviews_clean. >16k	40	4d 20h	31.7 MB	0	0
reviews_clean dataset's topic model	40	4d 20h	31.7 MB	0	0

Show 10 topic models

1 to 10 of 66 topic models

1 2 3 4 5

Figure 12.2: Stop topic model creation from pop up menu

Note: if you stop the topic model during its creation, you will not be able to resume the same task. If you want to create the same topic model, you will have to start a new task.

CHAPTER 13

Deleting Topic Models

You can delete your topic models by clicking the DELETE TOPIC MODEL option in the **1-click action menu** from the topic model view. (See Figure 13.1.)

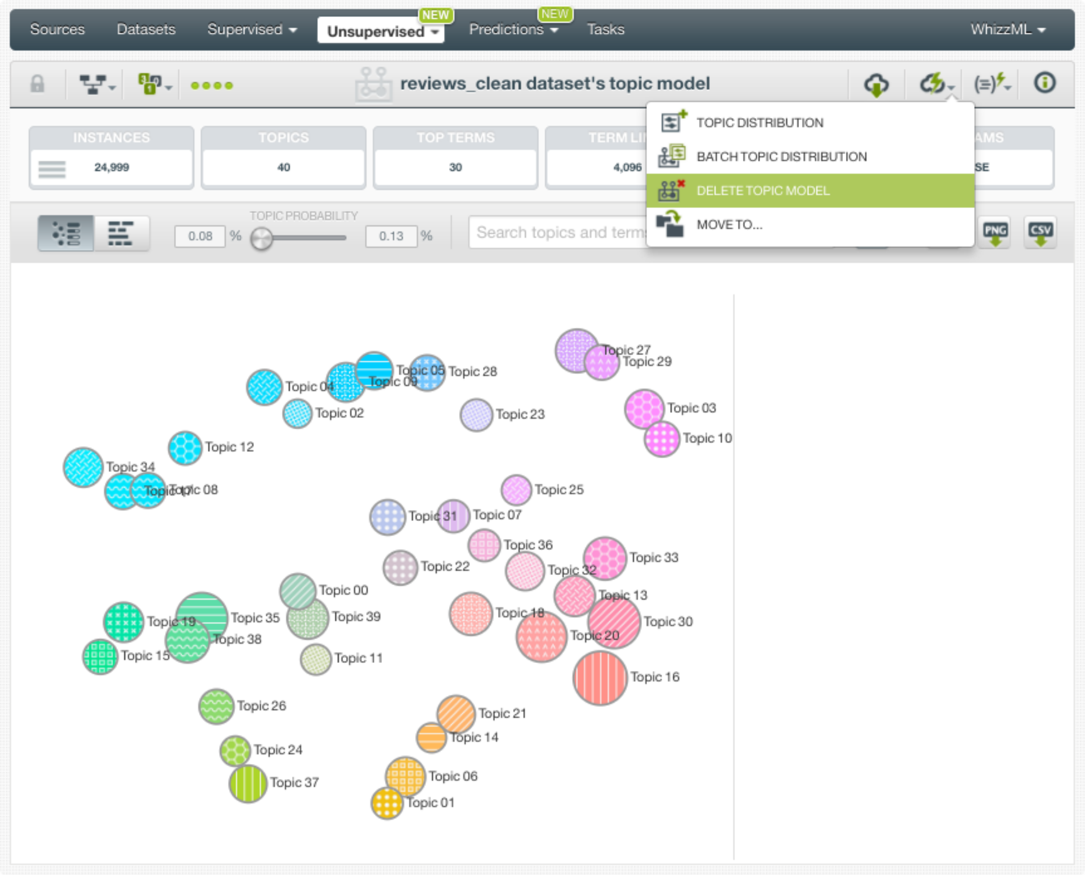
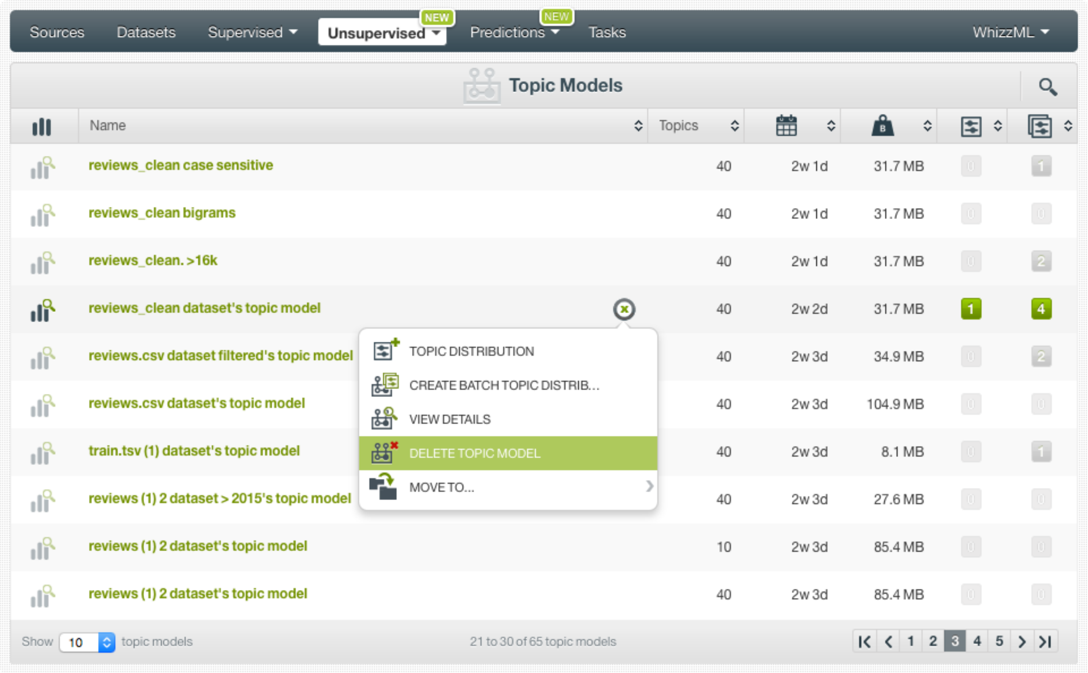


Figure 13.1: Delete topic models from 1-click menu

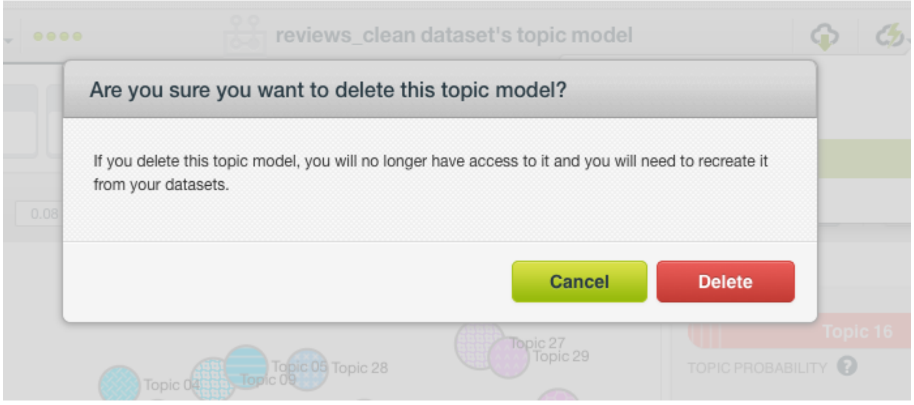
Alternatively, click the DELETE TOPIC MODEL in the **pop up menu** from the topic model list view. (See Figure 13.2.)



Name	Topics				
reviews_clean case sensitive	40	2w 1d	31.7 MB	0	1
reviews_clean bigrams	40	2w 1d	31.7 MB	0	0
reviews_clean. >16k	40	2w 1d	31.7 MB	0	2
reviews_clean dataset's topic model	40	2w 2d	31.7 MB	1	4
reviews.csv dataset filtered's topic model	40	2w 3d	34.9 MB	0	2
reviews.csv dataset's topic model	40	2w 3d	104.9 MB	0	0
train.tsv (1) dataset's topic model	40	2w 3d	8.1 MB	0	1
reviews (1) 2 dataset > 2015's topic model	40	2w 3d	27.6 MB	0	0
reviews (1) 2 dataset's topic model	10	2w 3d	85.4 MB	0	0
reviews (1) 2 dataset's topic model	40	2w 3d	85.4 MB	0	0

Figure 13.2: Delete topic models from pop up menu

A modal window will be displayed asking you for confirmation. After a topic model is deleted, it is permanently deleted, and there is no way you (or even the IT folks at BigML) can retrieve it.



Are you sure you want to delete this topic model?

If you delete this topic model, you will no longer have access to it and you will need to recreate it from your datasets.

Figure 13.3: Confirmation message to delete a topic model

CHAPTER 14

Takeaways

This document covered topic models in detail. We conclude it with a list of key points:

- Topic modeling is an **unsupervised** learning method used to discover the relevant topics in a collection of documents.
- BigML topic models use an optimized implementation of the **Latent Dirichlet Allocation** algorithm, one of the best-known probabilistic methods for topic modeling.
- BigML topic models only support **text fields** as inputs, the rest of fields will be ignored by the model.
- To create topic models you just need an existing **dataset** containing at least one text field. Then topic models can be used to make a single Topic Distribution or a Batch Topic Distribution. (See Figure 14.1.)
- You can use the **1-click option** to create your topic model or you can **configure** the several parameters provided by BigML before.
- When the topic model has been created, you get two different views: the topic chart and the term chart.
- The topic chart maps your topics as circles so you can get an overview of the topic importances and their relationships in the dataset.
- The term chart allows you to get a general view of your topic terms and their importances.
- You can use your topic model to **predict** the topic distributions over a single new instance or multiple instances simultaneously.
- You can create, configure, update, and use your topic models programmatically via the **BigML API and bindings**.
- You can download your topic models to **locally** calculate the topic probabilities for your new instances.
- You can add **descriptive information** to your topic models.
- You can **move** your topic models between projects.
- You can **share** your topic models with other people using the secret link or embedding them into your own applications.
- You can **stop** your topic models creation by deleting them.
- You can permanently **delete** your existing topic models.

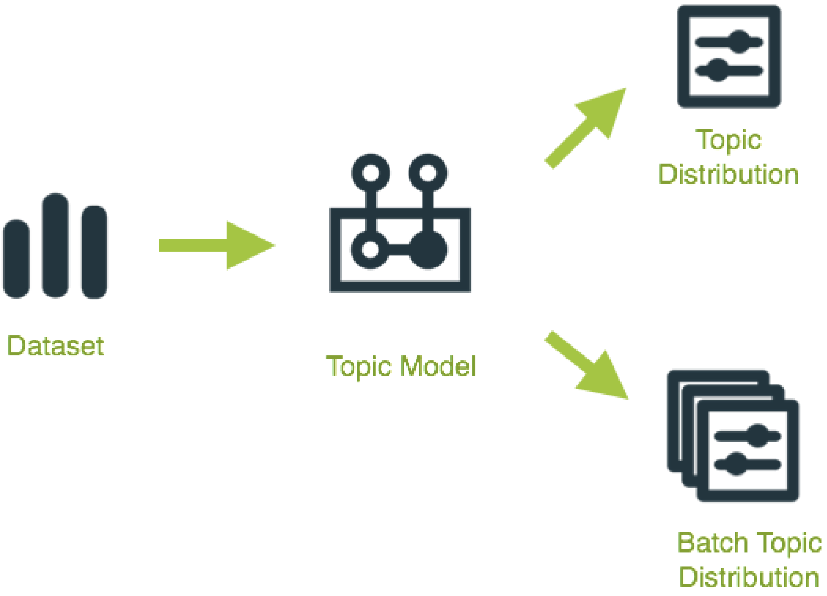


Figure 14.1: Topic models workflow

List of Figures

List of Tables

Glossary

Dashboard The BigML web-based interface that helps you privately navigate, visualize, and interact with your modeling resources. ii, 1

Field an attribute of each instance in your data. Also called "feature", "covariate", or "predictor". Each field is associated with a type (numeric, categorical, text, items, or date-time). 3

Instances the data points that represent the entity you want to model, also known as observations or examples. They are usually the rows in your data with a value (potentially missing) for each field that describes the entity. 3

Local predictions the predictions made in your local environment, faster, at no cost, by downloading your model. 55

Project an abstract resource that helps you group related BigML resources together. 2, 33, 64

Source the BigML resource that represents the data source to which you wish to apply Machine Learning. A data source stores an arbitrarily-large collection of instances. A BigML source helps you ensure that your data is parsed correctly. The BigML preferred format for data sources is tabular data in which each row is used to represent one of the instances, and each column is used to represent a field of each instance. 4, 9

Supervised learning a type of Machine Learning problem in which each instance of the data has a label. The label for each instance is provided in the training data, and a supervised Machine Learning algorithm learns a function or model that will predict the label given all other features in the data. The function can then be applied to data unseen during training to predict the label for unlabeled instances. ii

Topic the output of a topic model. Each topic is a distribution over terms that are thematically related. Each term has a different probability within a topic: the higher the probability, the more relevant is a term for that topic. ii, 1, 3

Topic Model an unsupervised Machine Learning task which identifies the relevant topics in the dataset text fields. Topic models in BigML are an optimized implementation of the Latent Dirichlet Allocation algorithm, a probabilistic method to find topics in large archive of documents. 1, 3

Topic Distribution a topic distribution is created using a topic model and the new text (input data) for which you wish to obtain the topic probabilities. ii, 55

Unsupervised learning a type of Machine Learning problem in which the objective is not to learn a predictor, and thus does not require each instance to be labeled. Typically, unsupervised learning algorithms infer some summarizing structure over the dataset, such as a clustering or a set of association rules. ii, 1, 3

References

[1] The BigML Team. *Anomaly Detection with the BigML Dashboard*. Tech. rep. BigML, Inc., Jan. 2016.

[2] The BigML Team. *Association Discovery with the BigML Dashboard*. Tech. rep. BigML, Inc., Dec. 2015.

[3] The BigML Team. *Classification and Regression with the BigML Dashboard*. Tech. rep. BigML, Inc., May 2016.

[4] The BigML Team. *Cluster Analysis with the BigML Dashboard*. Tech. rep. BigML, Inc., May 2016.

[5] The BigML Team. *Datasets with the BigML Dashboard*. Tech. rep. BigML, Inc., Jan. 2016.

[6] The BigML Team. *Sources with the BigML Dashboard*. Tech. rep. BigML, Inc., Jan. 2016.

[7] The BigML Team. *Time Series with the BigML Dashboard*. Tech. rep. BigML, Inc., July 2017.

bigml®