Sentiment Analysis

Predict sentiment from text.

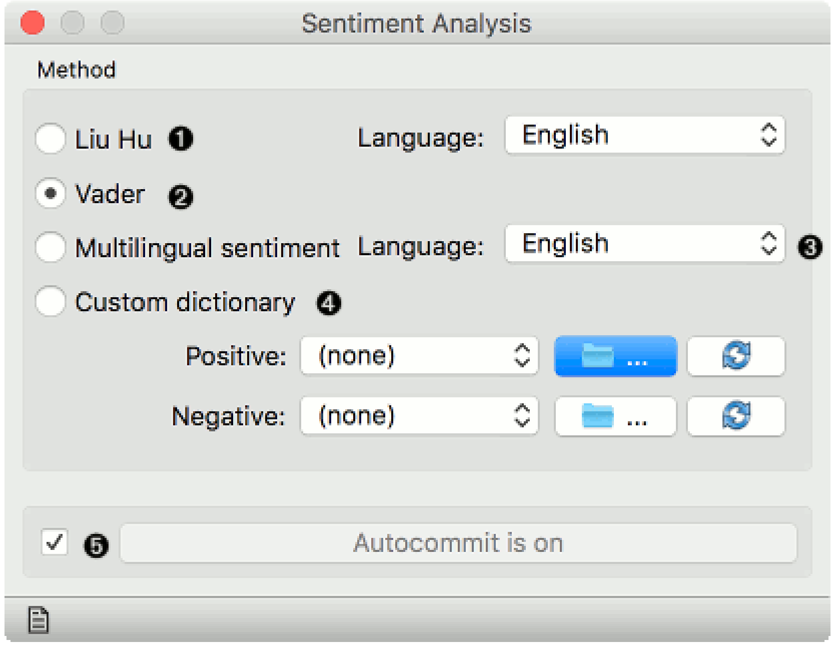
Inputs

- Corpus: A collection of documents.

Outputs

- Corpus: A corpus with information on the sentiment of each document.

Sentiment Analysis predicts sentiment for each document in a corpus. It uses Liu & Hu and Vader sentiment modules from NLTK and multilingual sentiment lexicons from the Data Science Lab. All of them are lexicon-based. For Liu & Hu, you can choose English or Slovenian version. Vader works only on English. Multilingual sentiment supports several languages, which are listed at the bottom of this page.



Sentiment Analysis

Method

○ Liu Hu ❶ Language: English

● Vader ❷

○ Multilingual sentiment Language: English ❸

○ Custom dictionary ❹

Positive: (none)

Negative: (none)

☑ ❺ Autocommit is on

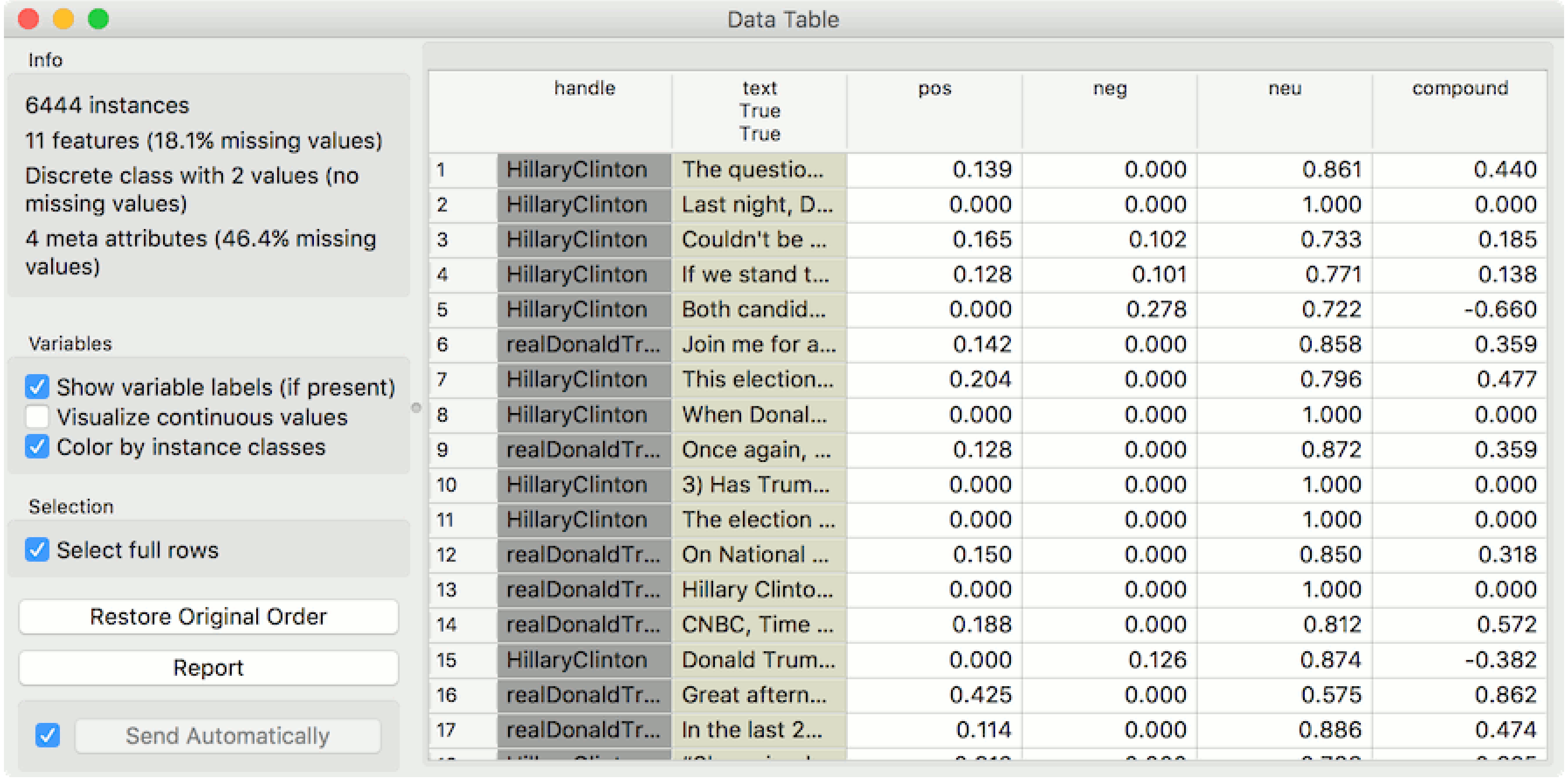
1. *Liu Hu*: lexicon-based sentiment analysis (supports English and Slovenian). The final score is the difference between the sum of positive and sum of negative words, normalized by the length of the document and multiplied by a 100. The final score reflects the percentage of sentiment difference in the document.
2. *Vader*: lexicon- and rule-based sentiment analysis
3. *Multilingual sentiment*: lexicon-based sentiment analysis for several languages

4. *Custom dictionary*: add you own positive and negative sentiment dictionaries. Accepted source type is .txt file with each word in its own line. The final score is computed in the same way as Liu Hu.
5. If *Auto commit is on*, sentiment-tagged corpus is communicated automatically. Alternatively press *Commit*.

Example

Sentiment Analysis can be used for constructing additional features with sentiment prediction from corpus. First, we load *Election-2016-tweets.tab* in Corpus. Then we connect **Corpus** to **Sentiment Analysis**. The widget will append 4 new features for Vader method: positive score, negative score, neutral score and compound (combined score).

We can observe new features in a **Data Table**, where we sorted the *compound* by score. Compound represents the total sentiment of a tweet, where -1 is the most negative and 1 the most positive.



Info

6444 instances

11 features (18.1% missing values)

Discrete class with 2 values (no missing values)

4 meta attributes (46.4% missing values)

Variables

☑ Show variable labels (if present)

☐ Visualize continuous values

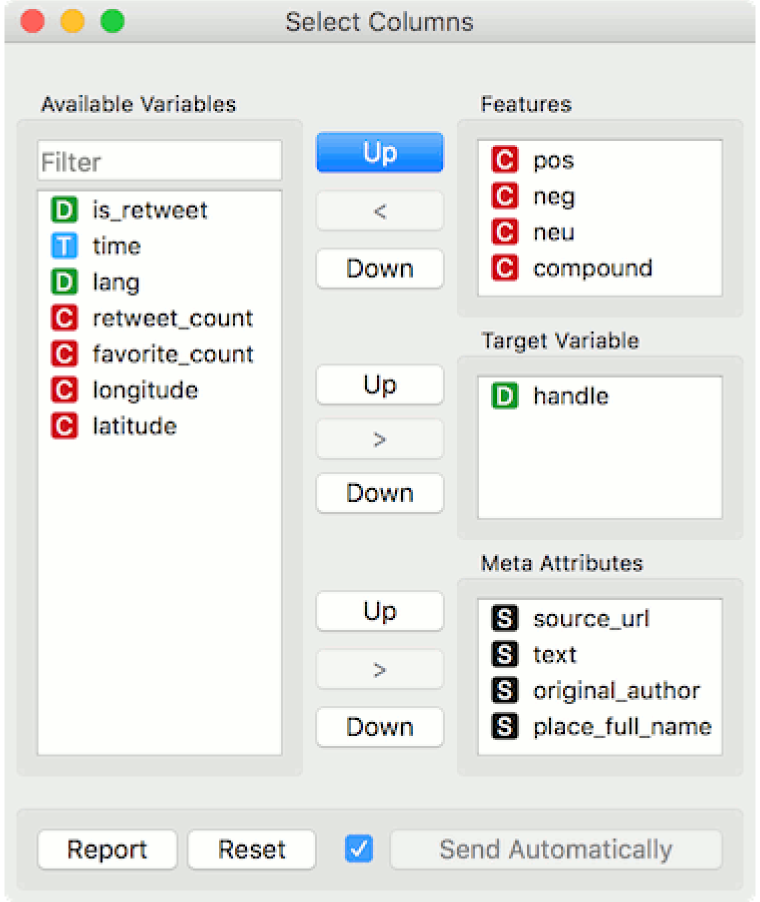
☑ Color by instance classes

Selection

☑ Select full rows

	handle	text	pos	neg	neu	compound
		True				
		True				
1	HillaryClinton	The questio...	0.139	0.000	0.861	0.440
2	HillaryClinton	Last night, D...	0.000	0.000	1.000	0.000
3	HillaryClinton	Couldn't be ...	0.165	0.102	0.733	0.185
4	HillaryClinton	If we stand t...	0.128	0.101	0.771	0.138
5	HillaryClinton	Both candid...	0.000	0.278	0.722	-0.660
6	realDonaldTr...	Join me for a...	0.142	0.000	0.858	0.359
7	HillaryClinton	This election...	0.204	0.000	0.796	0.477
8	HillaryClinton	When Donal...	0.000	0.000	1.000	0.000
9	realDonaldTr...	Once again, ...	0.128	0.000	0.872	0.359
10	HillaryClinton	3) Has Trum...	0.000	0.000	1.000	0.000
11	HillaryClinton	The election ...	0.000	0.000	1.000	0.000
12	realDonaldTr...	On National ...	0.150	0.000	0.850	0.318
13	realDonaldTr...	Hillary Clinto...	0.000	0.000	1.000	0.000
14	realDonaldTr...	CNBC, Time ...	0.188	0.000	0.812	0.572
15	HillaryClinton	Donald Trum...	0.000	0.126	0.874	-0.382
16	realDonaldTr...	Great aftern...	0.425	0.000	0.575	0.862
17	realDonaldTr...	In the last 2...	0.114	0.000	0.886	0.474

Now let us visualize the data. We have some features we are currently not interested in, so we will remove them with **Select Columns**.



Select Columns

Available Variables

Filter

D is_retweet
T time
D lang
C retweet_count
C favorite_count
C longitude
C latitude

Features

C pos
C neg
C neu
C compound

Target Variable

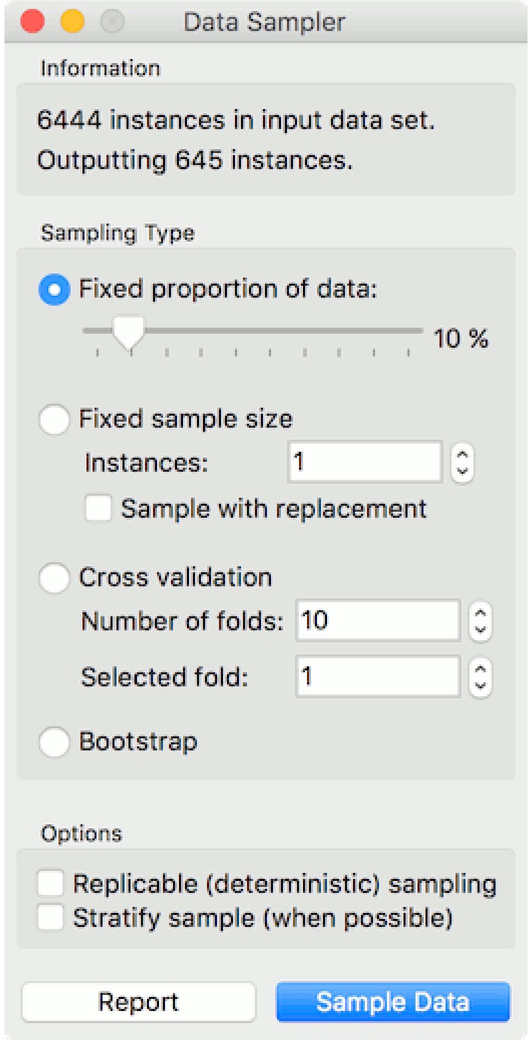
D handle

Meta Attributes

S source_url
S text
S original_author
S place_full_name

☑ Send Automatically

Then we will make our corpus a little smaller, so it will be easier to visualize. Pass the data to **Data Sampler** and retain a random 10% of the tweets.



6444 instances in input data set.
Outputting 645 instances.

● Fixed proportion of data: 10 %

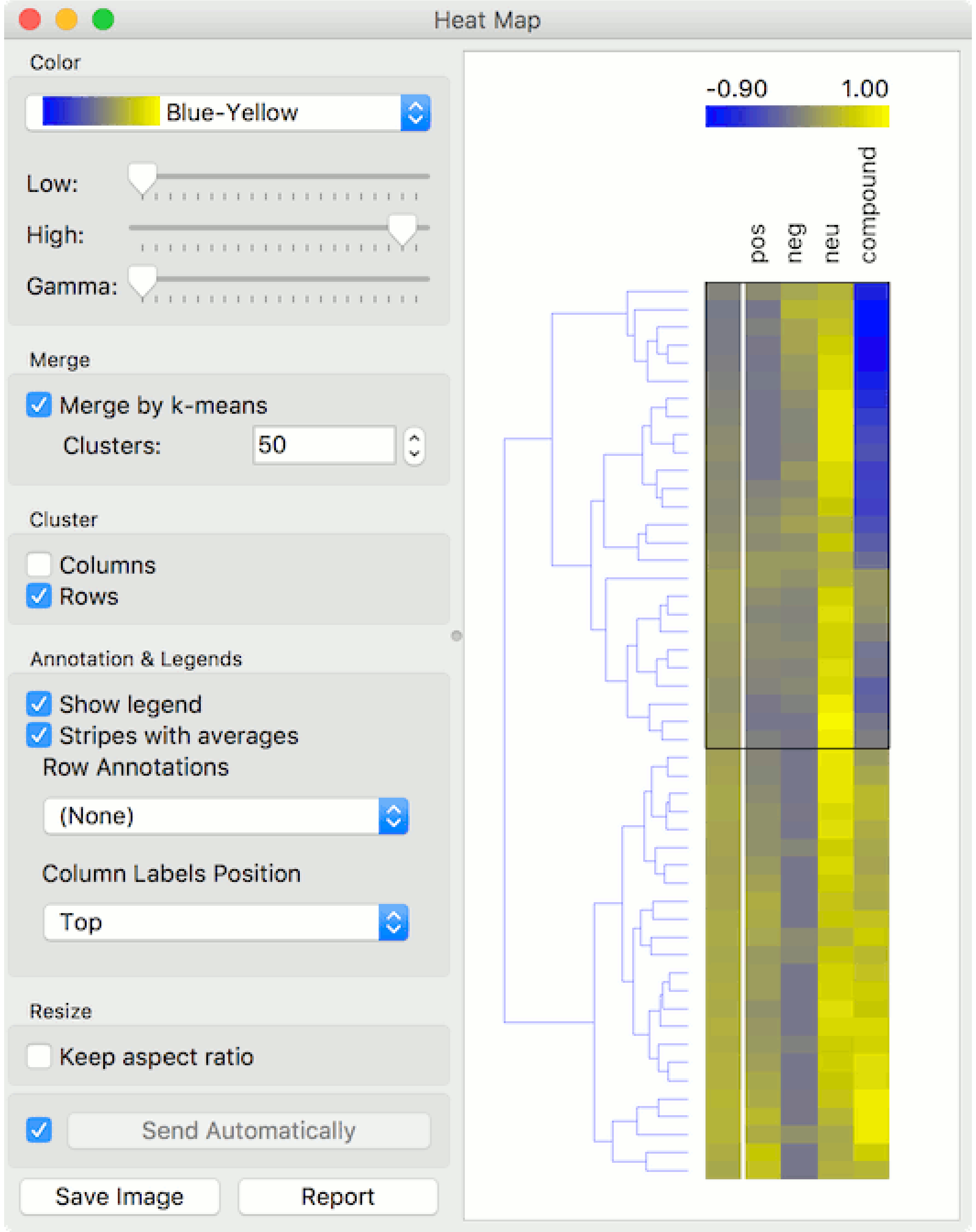
○ Fixed sample size
Instances: 1
□ Sample with replacement

○ Cross validation
Number of folds: 10
Selected fold: 1

○ Bootstrap

□ Replicable (deterministic) sampling
□ Stratify sample (when possible)

Now pass the filtered corpus to **Heat Map**. Use *Merge by k-means* to merge tweets with the same polarity into one line. Then use *Cluster* by *rows* to create a clustered visualization where similar tweets are grouped together. Click on a cluster to select a group of tweets - we selected the negative cluster.



Heat Map

Color

Blue-Yellow

Low:

High:

Gamma:

Merge

☑ Merge by k-means

Clusters: 50

Cluster

□ Columns

☑ Rows

Annotation & Legends

☑ Show legend

☑ Stripes with averages

Row Annotations

(None)

Column Labels Position

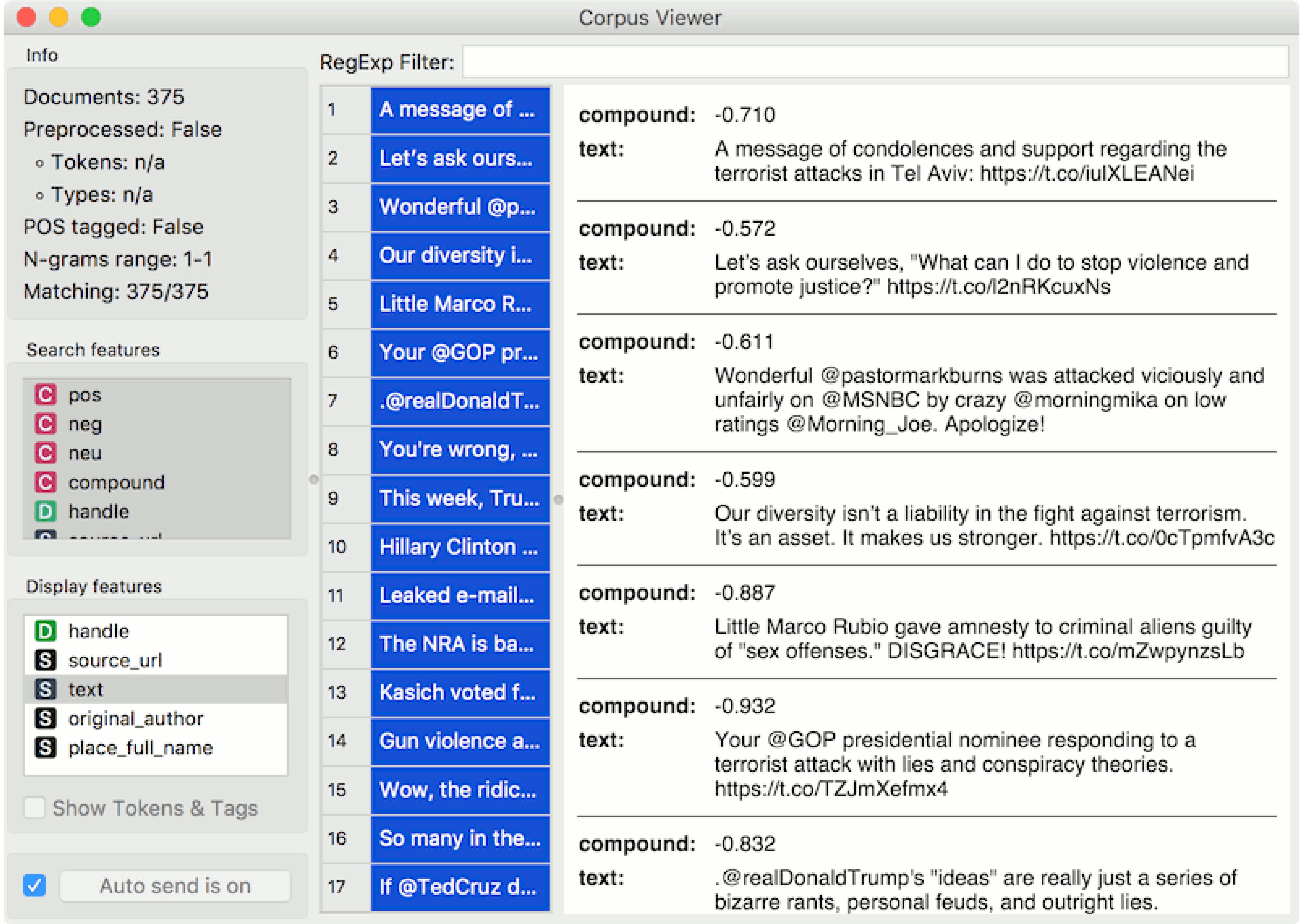
Top

Resize

□ Keep aspect ratio

☑ Send Automatically

To observe the selected subset, pass the tweets to Corpus Viewer.



Documents: 375
Preprocessed: False
- Tokens: n/a
- Types: n/a

POS tagged: False
N-grams range: 1-1
Matching: 375/375

compound: -0.710
text: A message of condolences and support regarding the terrorist attacks in Tel Aviv: https://t.co/iulXLEANei

compound: -0.572
text: Let's ask ourselves, "What can I do to stop violence and promote justice?" https://t.co/l2nRKcuxNs

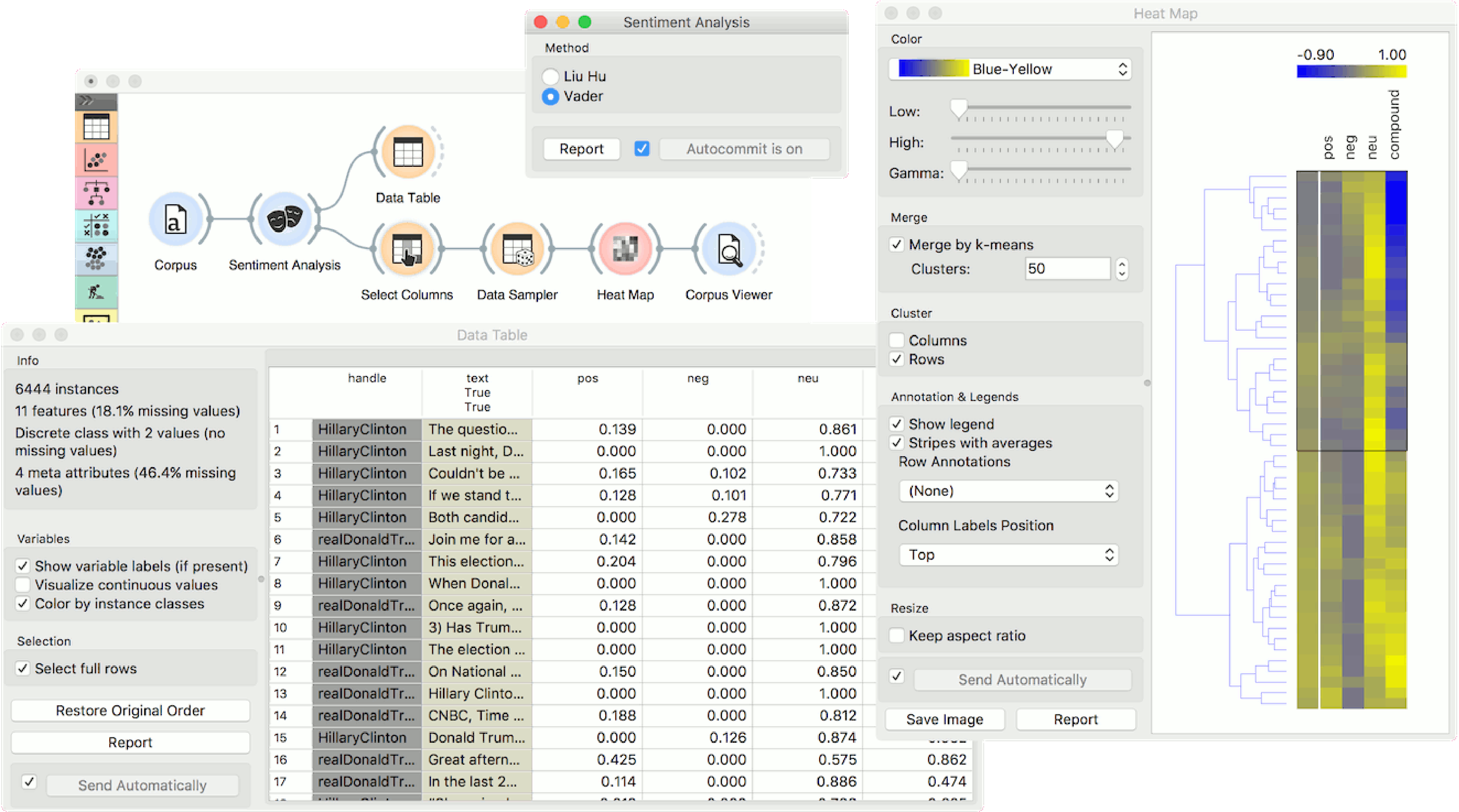
compound: -0.611
text: Wonderful @pastormarkburns was attacked viciously and unfairly on @MSNBC by crazy @morningmika on low ratings @Morning_Joe. Apologize!

compound: -0.599
text: Our diversity isn't a liability in the fight against terrorism. It's an asset. It makes us stronger. https://t.co/0cTpmfvA3c

compound: -0.887
text: Little Marco Rubio gave amnesty to criminal aliens guilty of "sex offenses." DISGRACE! https://t.co/mZwpynzsLb

compound: -0.932
text: Your @GOP presidential nominee responding to a terrorist attack with lies and conspiracy theories. https://t.co/TZJmXefmx4

compound: -0.832
text: .@realDonaldTrump's "ideas" are really just a series of bizarre rants, personal feuds, and outright lies.



Sentiment Analysis

Method

○ Liu Hu
● Vader

Report ☑ Autocommit is on

Heat Map

Color: Blue-Yellow

Low:
High:
Gamma:

Merge

☑ Merge by k-means
Clusters: 50

Cluster

□ Columns
☑ Rows

Annotation & Legends

☑ Show legend
☑ Stripes with averages
Row Annotations: (None)
Column Labels Position: Top

Resize

□ Keep aspect ratio

☑ Send Automatically

Save Image Report

Data Table

Info

6444 instances
11 features (18.1% missing values)
Discrete class with 2 values (no missing values)
4 meta attributes (46.4% missing values)

Variables

☑ Show variable labels (if present)
□ Visualize continuous values
☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

Report

☑ Send Automatically

	handle	text True True	pos	neg	neu	
1	HillaryClinton	The questio...	0.139	0.000	0.861	
2	HillaryClinton	Last night, D...	0.000	0.000	1.000	
3	HillaryClinton	Couldn't be ...	0.165	0.102	0.733	
4	HillaryClinton	If we stand t...	0.128	0.101	0.771	
5	HillaryClinton	Both candid...	0.000	0.278	0.722	
6	realDonaldTr...	Join me for a...	0.142	0.000	0.858	
7	HillaryClinton	This election...	0.204	0.000	0.796	
8	HillaryClinton	When Donal...	0.000	0.000	1.000	
9	realDonaldTr...	Once again, ...	0.128	0.000	0.872	
10	HillaryClinton	3) Has Trum...	0.000	0.000	1.000	
11	HillaryClinton	The election ...	0.000	0.000	1.000	
12	realDonaldTr...	On National ...	0.150	0.000	0.850	
13	realDonaldTr...	Hillary Clinto...	0.000	0.000	1.000	
14	realDonaldTr...	CNBC, Time ...	0.188	0.000	0.812	
15	HillaryClinton	Donald Trum...	0.000	0.126	0.874	
16	realDonaldTr...	Great aftern...	0.425	0.000	0.575	0.862
17	realDonaldTr...	In the last 2...	0.114	0.000	0.886	0.474

References

Hutto, C.J. and E. E. Gilbert (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Eighth International Conference on Weblogs and Social Media (ICWSM-14). Ann Arbor, MI, June 2014.

Hu, Minqing and Bing Liu (2004). Mining opinion features in customer reviews. In Proceedings of AAAI Conference on Artificial Intelligence, vol. 4, pp. 755–760. Available online.

Kadunc, Klemen and Marko Robnik-Šikonja (2016). Analiza mnenj s pomočjo strojnega učenja in slovenskega leksikona sentimenta. Conference on Language Technologies & Digital Humanities, Ljubljana (in Slovene). Available online.

Multilingual Sentiment Languages

- Afrikaans
- Arabic
- Azerbaijani
- Belarusian
- Bosnian
- Bulgarian
- Catalan
- Chinese
- Chinese Characters
- Croatian
- Czech
- Danish
- Dutch
- English
- Estonian
- Farsi
- Finnish
- French
- Gaelic
- German
- Greek
- Hebrew
- Hindi
- Hungarian
- Indonesian
- Italian
- Japanese
- Korean
- Latin
- Latvian
- Lithuanian
- Macedonian
- Norwegian
- Norwegian Nynorsk

- Polish
- Portuguese
- Romanian
- Russian
- Serbian
- Slovak
- Slovene
- Spanish
- Swedish
- Turkish
- Ukrainian