Similarity Hashing

Computes documents hashes.

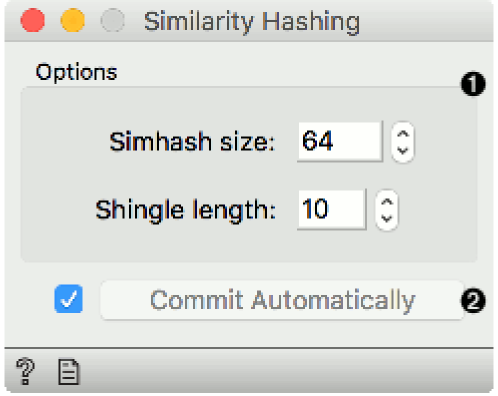
Inputs

- Corpus: A collection of documents.

Outputs

- Corpus: Corpus with simhash value as attributes.

Similarity Hashing is a widget that transforms documents into similarity vectors. The widget uses SimHash method from from Moses Charikar.



Similarity Hashing

Options

Simhash size: 64

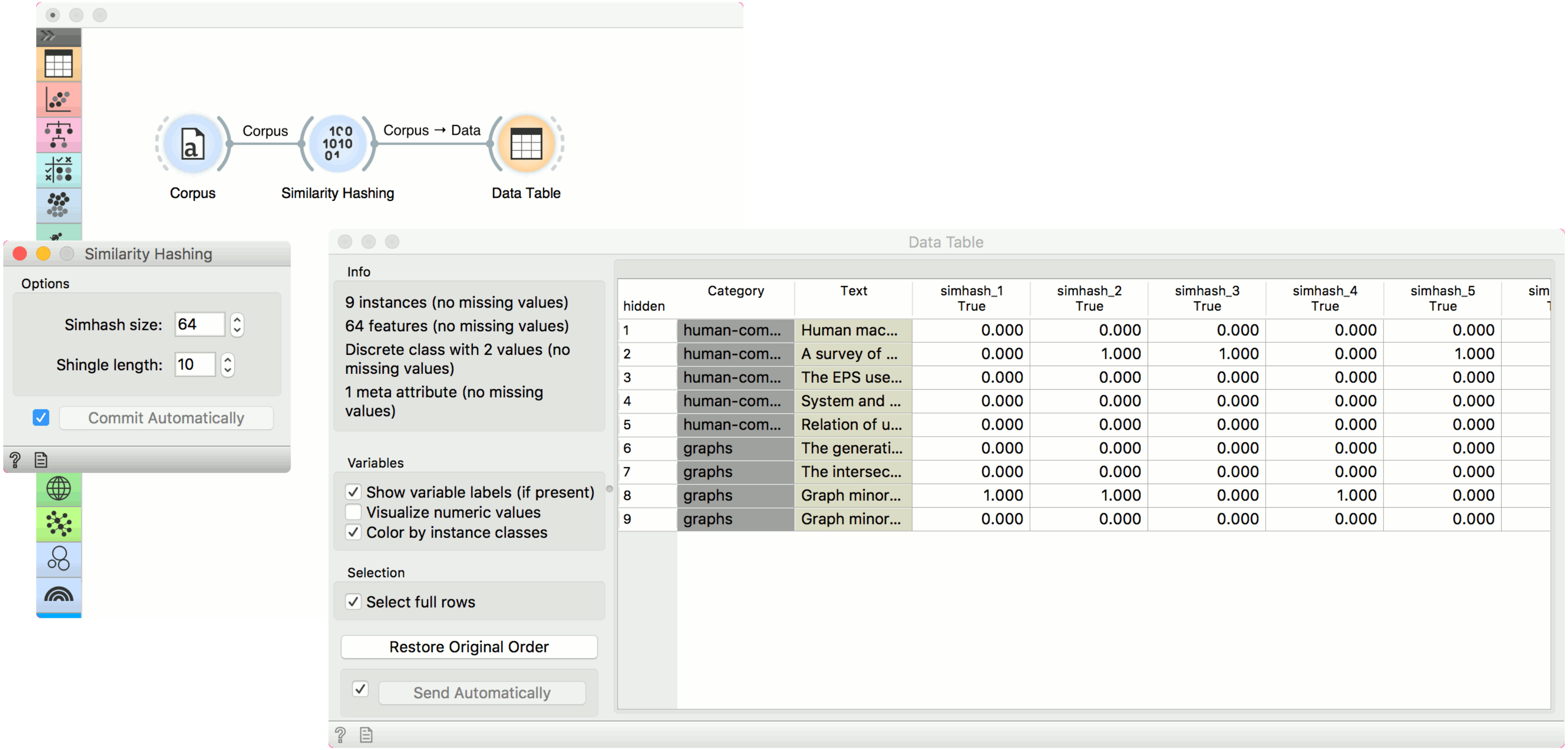
Shingle length: 10

☑ Commit Automatically

1. Set Simhash size (how many attributes will be on the output, corresponds to bits of information) and shingle length (how many tokens are used in a shingle).
2. *Commit Automatically* output the data automatically. Alternatively, press *Commit*.

Example

We will use *deerwester.tab* to find similar documents in this small corpus. Load the data with Corpus and pass it to **Similarity Hashing**. We will keep the default hash size and shingle length. We can observe what the widget outputs in a **Data Table**. There are 64 new attributes available, corresponding to the *Simhash size* parameter.



Similarity Hashing

Options

Simhash size: 64

Shingle length: 10

☑ Commit Automatically

Data Table

Info

9 instances (no missing values)

64 features (no missing values)

Discrete class with 2 values (no missing values)

1 meta attribute (no missing values)

Variables

☑ Show variable labels (if present)

☐ Visualize numeric values

☑ Color by instance classes

Selection

☑ Select full rows

Restore Original Order

☑ Send Automatically

hidden	Category	Text	simhash_1 True	simhash_2 True	simhash_3 True	simhash_4 True	simhash_5 True	sim T
1	human-com...	Human mac...	0.000	0.000	0.000	0.000	0.000	
2	human-com...	A survey of ...	0.000	1.000	1.000	0.000	1.000	
3	human-com...	The EPS use...	0.000	0.000	0.000	0.000	0.000	
4	human-com...	System and ...	0.000	0.000	0.000	0.000	0.000	
5	human-com...	Relation of u...	0.000	0.000	0.000	0.000	0.000	
6	graphs	The generati...	0.000	0.000	0.000	0.000	0.000	
7	graphs	The intersec...	0.000	0.000	0.000	0.000	0.000	
8	graphs	Graph minor...	1.000	1.000	0.000	1.000	0.000	
9	graphs	Graph minor...	0.000	0.000	0.000	0.000	0.000	

References

Charikar, M. (2002) Similarity estimation techniques from rounding algorithms. STOC ‘02 Proceedings of the thirty-fourth annual ACM symposium on Theory of computing, p. 380-