

Similarity Hashing

Computes documents hashes.

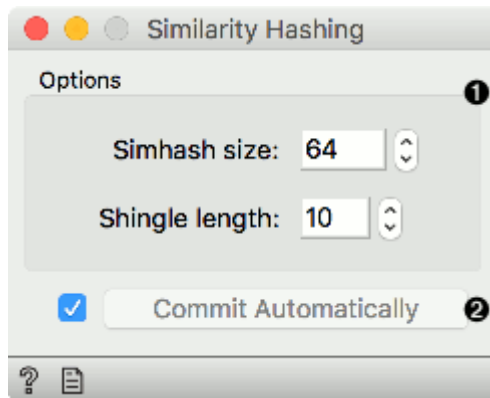
Inputs

- 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.



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.

The screenshot displays the Orange Data Mining software interface. At the top, a workflow is shown: a 'Corpus' widget (represented by a document icon with 'a') is connected to a 'Similarity Hashing' widget (represented by a circle with binary code), which is then connected to a 'Data Table' widget (represented by a grid icon). Below the workflow, the 'Similarity Hashing' widget's options are visible: 'Simhash size' is set to 64, 'Shingle length' is set to 10, and 'Commit Automatically' is checked. The 'Data Table' widget's 'Info' panel shows: '9 instances (no missing values)', '64 features (no missing values)', 'Discrete class with 2 values (no missing values)', and '1 meta attribute (no missing values)'. The 'Variables' panel has 'Show variable labels (if present)' checked, 'Visualize numeric values' unchecked, and 'Color by instance classes' checked. The 'Selection' panel has 'Select full rows' checked. The 'Data Table' widget's main area displays a table with 9 rows and 9 columns. The first column is 'hidden', the second is 'Category', the third is 'Text', and the remaining six are simhash_1 through simhash_6. The 'simhash_1' column contains the value 'True' for all rows. The 'Category' column has two values: 'human-com...' and 'graphs'. The 'Text' column contains various text snippets. The 'simhash' columns contain binary values (0.000 or 1.000).

hidden	Category	Text	simhash_1 True	simhash_2 True	simhash_3 True	simhash_4 True	simhash_5 True	simhash_6 True
1	human-com...	Human mac...	0.000	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	0.000
3	human-com...	The EPS use...	0.000	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	0.000
5	human-com...	Relation of u...	0.000	0.000	0.000	0.000	0.000	0.000
6	graphs	The generati...	0.000	0.000	0.000	0.000	0.000	0.000
7	graphs	The intersec...	0.000	0.000	0.000	0.000	0.000	0.000
8	graphs	Graph minor...	1.000	1.000	0.000	1.000	0.000	0.000
9	graphs	Graph minor...	0.000	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-