

# Bag of Words

Generates a bag of words from the input corpus.

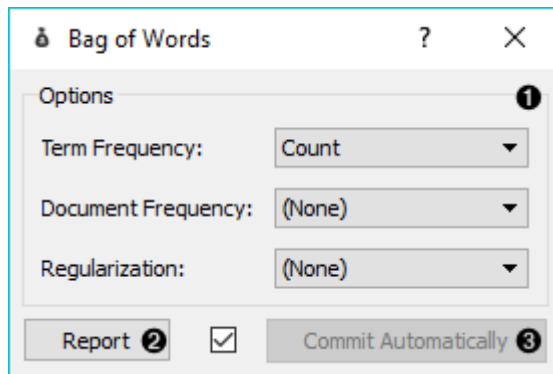
## Inputs

- Corpus: A collection of documents.

## Outputs

- Corpus: Corpus with bag of words features appended.

**Bag of Words** model creates a corpus with word counts for each data instance (document). The count can be either absolute, binary (contains or does not contain) or sublinear (logarithm of the term frequency). Bag of words model is required in combination with **Word Enrichment** and could be used for predictive modelling.



### 1. Parameters for **bag of words** model:

- Term Frequency:
  - Count: number of occurrences of a word in a document
  - Binary: word appears or does not appear in the document
  - Sublinear: logarithm of term frequency (count)
- Document Frequency:
  - (None)
  - IDF: **inverse document frequency**
  - **Smooth IDF**: adds one to document frequencies to prevent zero division.
- Regularization:
  - (None)
  - L1 (Sum of elements): normalizes vector length to sum of elements
  - L2 (Euclidean): normalizes vector length to sum of squares

### 2. Produce a report.

3. If *Commit Automatically* is on, changes are communicated automatically. Alternatively press *Commit*.

## Example

In the first example we will simply check how the bag of words model looks like. Load *book-excerpts.tab* with **Corpus** widget and connect it to **Bag of Words**. Here we kept the defaults - a simple count of term frequencies. Check what the **Bag of Words** outputs with **Data Table**. The final column in white represents term frequencies for each document.

untitled\*

File Edit View Widget Options Help

Corpus Bag of Words Data Table

**Bag of Words**

Options

Term Frequency: Count

Document Frequency: (None)

Regularization: (None)

Report ☒ Commit Automatically

**Data Table**

Info

140 instances  
10865 features (sparse, density 0.05%)  
Discrete class with 2 values (no missing values)  
1 meta attribute (no 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

| hidden | category | text                | {...}  |
|--------|----------|---------------------|--|
| 1      | children | the house Jim s...  | broke=1.000, by=4.000, trebly=1.000, basin=3.000, executed=1.000, picture=1.000, se...   |
| 2      | children | has lived rough...  | golden=1.000, carried=1.000, bar=2.000, confessions=1.000, air=1.000, again=5.000, r...  |
| 3      | children | Now boy he sai...   | gathering=1.000, letter=1.000, bring=1.000, resolved=1.000, payment=1.000, peculiar...   |
| 4      | children | thanks to you b...  | despair=1.000, thanks=1.000, finely=1.000, swift=1.000, terrors=1.000, rogues=1.000, ... |
| 5      | children | the empty ches...   | curiosity=1.000, drag=1.000, retreat=1.000, beyond=1.000, brief=1.000, cowardice=1....   |
| 6      | children | stood irresolute... | dance=3.000, furious=1.000, such=1.000, matter=1.000, fools=1.000, nearest=1.000, p...   |
| 7      | children | WE rode hard al...  | son=1.000, rascal=1.000, smoke=1.000, proud=1.000, hearty=1.000, villains=1.000, co...   |
| 8      | children | same as the tatt... | entry=1.000, roll=1.000, cache=1.000, blank=1.000, rank=1.000, manned=1.000, houn...     |
| 9      | children | IT was longer t...  | transparent=1.000, housekeeper=1.000, explored=1.000, fancies=1.000, plans=1.000, ...    |
| 10     | children | treasure Long J...  | dream=1.000, picked=1.000, telescope=1.000, substance=1.000, unearthed=1.000, ro...      |
| 11     | children | We are so grate...  | whatever=1.000, favor=1.000, therefore=1.000, beam=1.000, dismay=1.000, dwelt=1....      |
| 12     | children | I am told said t... | loudly=1.000, frock=1.000, bread=2.000, brook=1.000, around=1.000, grieve=1.000, g...    |
| 13     | children | to find the one ... | watched=2.000, chin=1.000, merrily=1.000, earnestly=1.000, stalks=1.000, stop=1.000...   |
| 14     | children | take away the p...  | unfriendly=1.000, nest=1.000, bites=1.000, truly=1.000, partv=1.000, lonesome=1.000...   |

In the second example we will try to predict document category. We are still using the *book-excerpts.tab* data set, which we sent through **Preprocess Text** with default parameters. Then we connected **Preprocess Text** to **Bag of Words** to obtain term frequencies by which we will compute the model.

The screenshot displays the Orange Data Mining software interface with a workflow for text classification. The workflow consists of the following widgets: Corpus, Preprocess Text, Bag of Words, SVM, Test & Score, Confusion Matrix, and Corpus Viewer.

**Bag of Words Widget Options:**

- Term Frequency: Count
- Document Frequency: IDF
- Regularization: (None)
- Buttons: Report, Commit Automatically

**Test & Score Widget Evaluation Results:**

| Method | AUC   | CA    | F1    | Precision | Recall |
|--------|-------|-------|-------|-----------|--------|
| SVM    | 0.971 | 0.971 | 0.971 | 1.000     | 0.943  |

**Confusion Matrix Widget:**

Learners: SVM

Show: Number of instances

Select: Select Correct, Select Misclassified, Clear Selection

Output: ☒ Predictions, ☐ Probabilities

☒ Send Automatically

Report

**Corpus Viewer Widget:**

Info:

- Documents: 4
- Preprocessed: False
- Tokens: n/a
- Types: n/a
- POS tagged: False
- N-grams range: 1-1
- Matching: 4/4

Search features:

- category
- text
- category(SVM)

Display features:

- category
- text
- category(SVM)

RegExp Filter:

| Document     | category | text  |
|--------------|----------|---|
| 1 Document 1 | children | thanks to you big hulking chicken-hearted men We'll have that chest open if we die for it And I'll thank you for that bag Mrs Crossley to bring back our lawful money in Of course I said I would go with my mother and of course they all cried out at our foolhardiness but even then not a man would go along with us All they would do was to give me a loaded pistol lest we were attacked and to promise to have horses ready saddled in case we were pursued on our return while one lad was to ride forward to the doctor's in search of armed assistance My heart was beating finely when we two set forth in the cold night upon this dangerous venture A full moon was beginning to rise and peered redly through the upper edges of the fog and this increased our haste for it was plain before we came forth again that all would be as bright as day and our departure exposed to the eyes of any watchers We slipped along the hedges noiseless and swift nor did we see or hear anything to increase our terrors till to our relief the door of the Admiral Benbow had closed behind us I slipped the bolt at once and we stood and panted for a moment in the dark alone in the house with the dead captain's body Then my mother got a candle in the bar and holding each other's hands we edged into the parlour. |
| 2 Document 2 |          |   |
| 3 Document 3 |          |   |
| 4 Document 4 |          |   |

Connect **Bag of Words** to **Test & Score** for predictive modelling. Connect **SVM** or any other classifier to **Test & Score** as well (both on the left side). **Test & Score** will now compute performance scores for each learner on the input. Here we got quite impressive results with SVM. Now we can check, where the model made a mistake.

Add **Confusion Matrix** to **Test & Score**. Confusion matrix displays correctly and incorrectly classified documents. *Select Misclassified* will output misclassified documents, which we can further inspect with **Corpus Viewer**.