

# Package: klsh (via r-universe)

September 5, 2024

**Type** Package

**Title** Blocking for Record Linkage

**Version** 0.1.0

**Depends** R (>= 3.0.2), blink, stats, utils, plyr

**Imports** Repp, stringi, SnowballC

**Suggests** knitr, ggplot2, rmarkdown

**VignetteBuilder** knitr

**Description** An implementation of the blocking algorithm KLSH in Steorts, Ventura, Sadinle, Fienberg (2014) <[DOI:10.1007/978-3-319-11257-2\\_20](https://doi.org/10.1007/978-3-319-11257-2_20)>, which is a k-means variant of locality sensitive hashing. The method is illustrated with examples and a vignette.

**Encoding** UTF-8

**LazyData** true

**License** GPL-3

**RoxygenNote** 7.1.1.9000

**Repository** <https://cleanzr.r-universe.dev>

**RemoteUrl** <https://github.com/cleanzr/klsh>

**RemoteRef** HEAD

**RemoteSha** de9921997ec1c86044a8648899fb883c352a78b1

## Contents

bag_of_word_ify . . . . .	2
bag_signatures . . . . .	2
block.ids.from.blocking . . . . .	3
calc_idf . . . . .	4
confusion.from.blocking . . . . .	4
klsh . . . . .	5
reduction.ratio . . . . .	6
reduction.ratio.from.blocking . . . . .	6

rproject_bags . . . . .	7
sacks_of_bags_of_words . . . . .	8
tokenify . . . . .	8

<b>Index</b>	<b>10</b>
--------------	-----------

---

bag_of_word_ify	<i>Function to convert a record into a bag of tokens with a fieldwise flag</i>
-----------------	--

---

### Description

Function to convert a record into a bag of tokens with a fieldwise flag

### Usage

```
bag_of_word_ify(record, k, fieldwise = FALSE)
```

### Arguments

record	String or record
k	Parameter k, which is the number of shingle, tokens, or grams to break the string into
fieldwise	Flag where the default setting to include the record as the entire string

### Value

Computes the bag of tokens for a string

### Examples

```
data(RLdata500)
data.500 <- RLdata500[-c(2,4)]
bag_of_word_ify(data.500[1,c(-2)],k=2)
bag_of_word_ify(data.500[300,c(-2)],k=2)
names(bag_of_word_ify(data.500[300,c(-2)],k=2))
```

---

bag_signatures	<i>Function that reduces a bag of words into a signature matrix using multiple random projections</i>
----------------	---

---

### Description

Function that reduces a bag of words into a signature matrix using multiple random projections

### Usage

```
bag_signatures(sack_of_bags, p, weighting_table)
```

### Arguments

sack\_of\_bags     Sack of bag of words  
p                 Number of random projections p  
weighting\_table     Weighting table (inverse document frequency)

### Value

Computes a signature matrix using multiple random projections and the inverse document frequency weights

### Examples

```
data(RLdata500)
data.500 <- RLdata500[-c(2,4)]
sack <- sacks_of_bags_of_words(data.500[1:3,c(-2)],k=2)
idf <- calc_idf(sack)
bag_signatures(sack, p=5, idf)
```

---

`block.ids.from.blocking`

*Returns the block ids associated with a blocking method.*

---

### Description

Returns the block ids associated with a blocking method.

### Usage

```
block.ids.from.blocking(blocking)
```

### Arguments

blocking             A list of the blocks.

### Value

A list of the blocks ids that corresponds to each block

### Examples

```
data("RLdata500")
klsh.blocks <- klsh(RLdata500, p=20, num.blocks=5, k=2)
block.ids.from.blocking(klsh.blocks)
```

---

calc_idf	<i>Function to calculate the inverse document frequency given a shingled bag of words</i>
----------	---

---

**Description**

Function to calculate the inverse document frequency given a shingled bag of words

**Usage**

```
calc_idf(sack_of_bags)
```

**Arguments**

sack\_of\_bags    Sack of bag of words

**Value**

Computes the inverse document frequency for a bag of words

**Examples**

```
data(RLdata500)
data.500 <- RLdata500[-c(2,4)]
sack <- sacks_of_bags_of_words(data.500[1:3,c(-2)],k=2)
(idf <- calc_idf(sack))
match(names(sack[[1]]), names(idf))
```

---

confusion.from.blocking	<i>Perform evaluations (recall) for blocking.</i>
-------------------------	---

---

**Description**

Perform evaluations (recall) for blocking.

**Usage**

```
confusion.from.blocking(blocking, true_ids, recall.only = FALSE)
```

**Arguments**

blocking        A list of the blocks  
true\_ids        The true identifiers for comparisons  
recall.only     Flag that when true only prints the recall, otherwise prints many evaluation metrics in a list

**Value**

A vector of that returns the recall and the precision

**Examples**

```
data("RLdata500")
klsh.blocks <- klsh(RLdata500, p=20, num.blocks=5, k=2)
confusion.from.blocking(klsh.blocks, identity.RLdata500)
confusion.from.blocking(klsh.blocks, identity.RLdata500, recall.only=TRUE)
```

---

klsh	<i>Function that reduces a bag of words into a signature matrix using multiple random projections</i>
------	---

---

**Description**

Function that reduces a bag of words into a signature matrix using multiple random projections

**Usage**

```
klsh(r.set, p, num.blocks, k, fieldwise = FALSE, quiet = TRUE)
```

**Arguments**

r.set	Set of records
p	Number of random projections
num.blocks	The total number of desired blocks
k	The total number of tokens
fieldwise	Flag with default FALSE
quiet	Flag to turn on printed progress, default to TRUE

**Value**

The blocks from performing KLSH

**Examples**

```
data(RLdata500)
data.500 <- RLdata500[-c(2,4)]
klsh.blocks <- klsh(data.500, p=20, num.blocks=5, k=2)
```

reduction.ratio      *Returns the reduction ratio associated with a blocking method*

---

**Description**

Returns the reduction ratio associated with a blocking method

**Usage**

```
reduction.ratio(block.labels)
```

**Arguments**

block.labels      A list of the blocks labels.

**Value**

The reduction ratio

**Examples**

```
data("RLdata500")
klsh.blocks <- klsh(RLdata500, p=20, num.blocks=5, k=2)
block.ids <- block.ids.from.blocking(klsh.blocks)
reduction.ratio(block.ids)
```

---

reduction.ratio.from.blocking  
*Returns the reduction ratio associated with a blocking method*

---

**Description**

Returns the reduction ratio associated with a blocking method

**Usage**

```
reduction.ratio.from.blocking(blocking)
```

**Arguments**

blocking          The actual blocks

**Value**

The reduction ratio

**Examples**

```
data("RLdata500")
klsh.blocks <- klsh(RLdata500, p=20, num.blocks=5, k=2)
reduction.ratio.from.blocking(klsh.blocks)
```

---

rproject_bags	<i>Function that generates unit random vectors and takes (weighted) projections onto the random unit vectors given a bag of words</i>
---------------	---

---

**Description**

Function that generates unit random vectors and takes (weighted) projections onto the random unit vectors given a bag of words

**Usage**

```
rproject_bags(sack_of_bags, weighting_table)
```

**Arguments**

sack\_of\_bags    Sack of bag of words  
weighting\_table    Weighting table (inverse document frequency)

**Value**

Computes the inverse document frequency for a bag of words

**Examples**

```
data(RLdata500)
data.500 <- RLdata500[-c(2,4)]
sack <- sacks_of_bags_of_words(data.500[1:3,c(-2)],k=2)
idf <- calc_idf(sack)
match(names(sack[[1]]), names(idf))
rproject_bags(sack, idf)
```

---

sacks\_of\_bags\_of\_words

*Function to convert all records into a bag of tokens*

---

### Description

Function to convert all records into a bag of tokens

### Usage

```
sacks_of_bags_of_words(r.set, k, fieldwise = FALSE)
```

### Arguments

r.set	Record set
k	Parameter k, which is the number of shingle, tokens, or grams to break the string into
fieldwise	Flag where the default setting to include the record as the entire string

### Value

Computes the bag of tokens for a record set

### Examples

```
data(RLdata500)
data.500 <- RLdata500[-c(2,4)]
sacks_of_bags_of_words(data.500[1:3,c(-2)],k=2)
```

---

tokenify

*Function to token a string into its k components*

---

### Description

Function to token a string into its k components

### Usage

```
tokenify(string, k)
```

### Arguments

string	A string or record
k	A parameter k, which is the number of shingle, tokens, or grams to break the string into



**Value**

Computes the tokenized or grammed version of a string

**Examples**

```
tokenify("Alexander",2)  
tokenify("Alexander Smith", 2)
```

# Index

bag\_of\_word\_ify, 2  
bag\_signatures, 2  
block.ids.from.blocking, 3  
  
calc\_idf, 4  
confusion.from.blocking, 4  
  
klsh, 5  
  
reduction.ratio, 6  
reduction.ratio.from.blocking, 6  
rproject\_bags, 7  
  
sacks\_of\_bags\_of\_words, 8  
  
tokenify, 8