Word association measure

Bernard Desgraupes and Sylvain Loiseau
<bernard.desgraupes@u-paris10.fr>, <sylvain.loiseau@univ-paris13.fr>

October 22, 2015

Abstract

Contents

```
> library(wam);
```

1 Indicator of word association

Each association measure return a numeric vector indicating, for each corresponding index in the arguments, the association strengh between the word under scrutiny and the subcorpus.

These association measures, unless otherwise stated in the help page of the function, are positive when the word is over-represented ("attracted"), and negative when the word is under-represented.

In absolute value, the more the word is over-representend or under-represented, the more the association measure givien is hight.

1.1 Arguments of the functions

All functions have the following signature: (N, n, K, k), where:

1. N is the total size of the corpora
2. n is the size of the subcorpora
3. K is the total frequency of the form under scrutiny in the corpora
4. k is the sub-frequency of the form under scrutinty in the subcorpora

This can be easily turn into the "contingency table" representation used in some presentation (according to Stefan Evert UCS documentation) :

word	$\neg word$	T	
subcorpus	O11	O12	R1
	E11	E12	
$\neg subcorpus$	O21	O22	R2
	E21	E22	
Totals	C1	C2	N

where :

- N = total words in corpus (or subcorpus or restriction, but they are not implemented yet)
- C1 = frequency of the collocate in the whole corpus
- C2 = frequency of words that aren't the collocate in the corpus
- R1 = total words in window
- R2 = total words outside of window
- O11 = how many of collocate there are in the window

- O12 = how many words other than the collocate there are in the window (calculated from row total)
- O21 = how many of collocate there are outside the window
- O22 = how many words other than the collocate there are outside the window
- E11 = expected values (proportion of collocate that would belong in window if collocate were spread evenly)
- E12 = " " (proportion of collocate that would belong outside window if collocate were spread evenly)
- E21 = " " (proportion of other words that would belong in window if collocate were spread evenly)
- E22 = " " (proportion of other words that would belong outside window if collocate were spread evenly)

Conversion from N, n, K, k notation :

	word	$\neg word$	T
subcorpus	k	$n-k$	n
$\neg subcorpus$	$K-k$	$N-K-(n-k)$	$N-n$
Totals	K	$N-K$	N

Conversion to N, n, K, k notation :

- $N = N$
- $n = O11 + O12$
- $K = O11 + O21$
- $k = O11$

We will use the arguments from the data set *robespierre*:

```
> data(robespierre)
> robespierre <- robespierre[1:5,]
> robespierre

      N    n    K   k      types parts
1 61449 8395 3173 464         de    D1
2 61449 2558  296  45     peuple    D2
3 61449 3920  207  35 republique    D3
4 61449 6903  165  30     ennemi    D4
5 61449 7896  153   6     patrie    D5

> attach(robespierre)
```

1.2 Recycling arguments

Arguments are recycled.

1.3 Log-likelihood

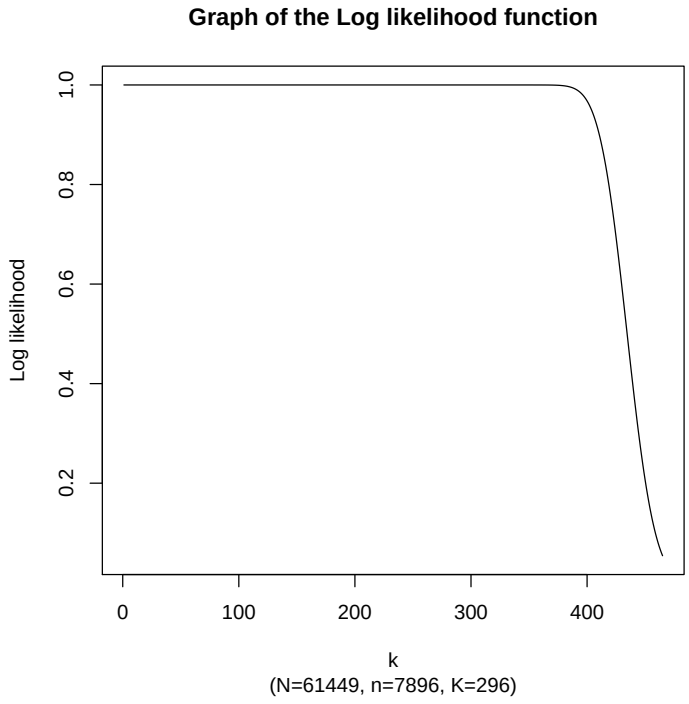
A given value of attraction between a form and a subcorpus:

```
> wam.loglikelihood(N, n, K, k);
```

```
[1] 5.425296e-02 4.494012e-14 8.616378e-08 4.298495e-03 9.999284e-01
```

Graph of the function:

```
> plot(wam.loglikelihood(N[1], n[1], K[1], 0:k[1]),
+      type="l", xlab="k", ylab="Log likelihood",
+           main="Graph of the Log likelihood function",
+           sub="(N=61449, n=7896, K=296)")
```



1.4 Specificities

This is an implementation of the indicator that has been proposed by Lafon in "Sur la variabilité de la fréquence des formes dans un corpus", *Mots*, 1, 1980, 127–165 (`http://www.persee.fr/web/revues/home/prescript/article/mots_0243-6450_1980_num_1_1_1008`).

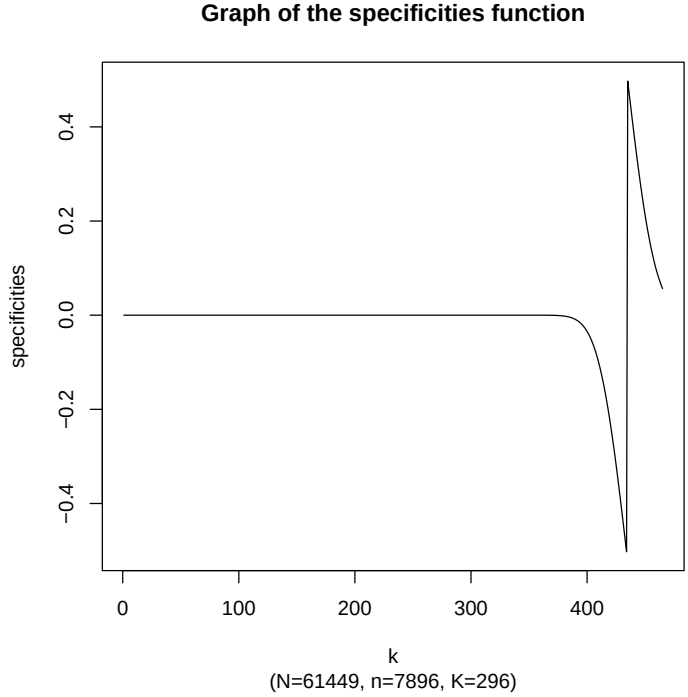
A given value of attraction between a form and a subcorpus:

```
> wam.specificities(N, n, K, k);
```

```
[1]  2.184893 29.638212 15.280475  4.544573 -8.037906
```

Graph of the function:

```
> plot(wam.specificities(N[1], n[1], K[1], 0:k[1], method="base"), type="l", xlab="k",
```



These exemple is the lexical frequency of the form *peuple* (French for people) in three public discourses by Robespierre in a corpus of 10 discourses containing $N = 61449$ occurrences in total (Lafon 1980) :

Discours	N	n	K	k
4	61449	6903	296	14
5	61449	7896	296	53
8	61449	2063	296	16

For each line we can compute the expected frequency of the form ($K \times n/N$) and mark + if the form is more frequent than expected or − otherwise.

Discours	N	n	K	k	expected	$k > expected$
4	61449	6903	296	14	32.80	−
5	61449	7896	296	53	37.52	+
8	61449	2063	296	16	9.80	+

The form *peuple* is less frequent in the fourth discourse than expected. On the contrary, *peuple* is more frequent than expected in the fifth and eigth discourses.

If the observed frequency is less than the expected frequency, we compute the sum of the probability for a frequency lesser or equal to the observed frequency ($Prob(X \leq k)$). If the observed frequency is greater than the expected frequency, we compute the sum of the probability for a frequency greater to the observed frequency ($Prob(X > k)$) (Lafon 1980 : 152).

In both cases, the more unexpected is the frequency, the smaller is the indicator.

Discours	N	n	K	k	expected	$k > expected$	cumulative extreme probability
4	61449	6903	296	14	32.80	−	0.0000669371
5	61449	7896	296	53	37.52	+	0.0077234888
8	61449	2063	296	16	9.80	+	0.0433282491

According to this indicator, the second case is more "surprising" than the third or, in other terms, *peuple* is more attracted by, or specific to the second discourse than to the third.

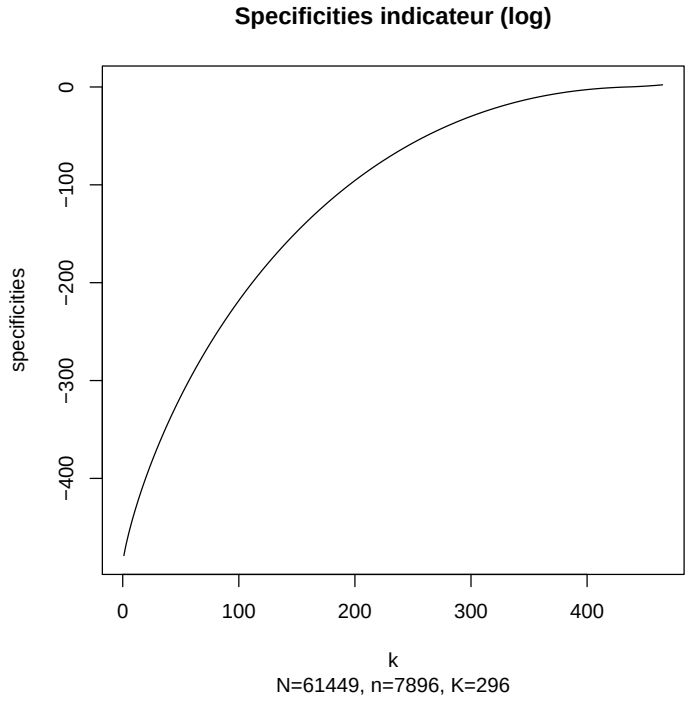
According to relative frequency, one could conclude the other way around: $53/7896 = 0.0067 < 16/2063 = 0.0078$ (cf. Lafon 1980 : 152).

For the fifth discourse above ($N = 61449$, $n = 7896$), the possible frequencies of *peuple* range from 0 to 296 (if all occurrences of *peuple* where in this discours). Here is the corresponding values for the specificities indicator:

1.4.1 Log of specificities

The log of the probability; with "-" sign if specificity is negative and "+" if it is positive.

```
> plot(wam.specificities(N[1], n[1], K[1], 0:k[1], method="log"), type="l", xlab="k", y
```



2 High level interface

The function *wam* provide a more high level interface to the actual functions.

This function allows for computing several indicator at the same time.

It takes also as argument the subcorpus name and lexical types.

```
> attach(robespierre)
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"), parts, types)
```

function allows for retrieving the basic information:

```
> attach(robespierre)
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"), parts, types)
```

A print implementation allows for an easy to use reading of the results:

```
> attach(robespierre)
> wam.res <- wam(N, n, K, k, measure=c("loglikelihood", "specificities"), parts, types)
> wam.res
```

```
Printing association measure for 5 part(s); from: 1 to: 100
Corpus size: 61449
Sorted by: loglikelihood
-------------------------------------------------------------------------------
word                 | sub freq | tot freq |   loglikelihood |   specificities
-------------------------------------------------------------------------------
.......................................................................
Part name: de
Part size: 8395 tokens.
Positive specificities printed: 1
Negative specificities printed: 0
D1                   |      464 |     3173 |            0.05 |            2.18
.......................................................................
Part name: ennemi
Part size: 6903 tokens.
Positive specificities printed: 1
Negative specificities printed: 0
D4                   |       30 |      165 |            0.00 |            4.54
.......................................................................
Part name: patrie
Part size: 7896 tokens.
Positive specificities printed: 1
Negative specificities printed: 0
D5                   |        6 |      153 |            1.00 |           -8.04
.......................................................................
Part name: peuple
Part size: 2558 tokens.
Positive specificities printed: 1
Negative specificities printed: 0
D2                   |       45 |      296 |            0.00 |           29.64
.......................................................................
Part name: republique
Part size: 3920 tokens.
Positive specificities printed: 1
Negative specificities printed: 0
D3                   |       35 |      207 |            0.00 |           15.28
```

3 Bibliographie

Dunning, T. 1993. "Accurate methods for the statistics of surprise and coincidence." In: Computational Linguistics. 19(1). Pp 61–74.

http://acl.ldc.upenn.edu/J/J93/J93-1003.pdf

Hofland, K. and Johanssen, S. 1989. Frequency analysis of English vocabulary and grammar, based on the LOB corpus. Oxford: Clarendon.

Kilgarriff, A. 1996. "Which words are particularly characteristic of a text? A survey of statistical approaches." In: Proceedings, ALLC-ACH '96. Bergen, Norway.

http://www.cse.iitb.ac.in/~shwetaghonge/prec_recall.pdf

Chaudhari, D. L., Damani, O. P. & Laxman, S. 2011. "Lexical co-occurrence, statistical significance, and word association". In: Conference on Empirical Methods in Natural Language Processing (Edinburgh, Scotland, UK, July 27-31). pp. 1058-68

http://www.aclweb.org/anthology-new/D/D11/D11-1098.pdf