Additional information regarding keyword extraction using loglikelihood measure

Using loglikelihood is a common approach to identify keywords in a study corpus (see the presentation by Gisle Andersen, lecture 10). The basic idea is to compare observed and expected frequencies of words in two corpora, a reference corpus and the study corpus.

	Corpus.A	Corpus.B	Total.row
Freq of word	a	b	a+b
Freq of other words	c-a	d-b	c+d-a-b
Total column	c	d	c+d

Expected values E1 and E2 are calculated as follows:

$E1 = c * (a + b)/(c + d)$

$E2 = d * (a + b)/(c + d)$

LL is calculated by:

$2 * ((a * log(a/E1)) + (b * log(b/E2)))$

Example 1: the word 'quarter'

We use the British National Corpus as a reference corpus and a corpus of earning call transcripts as our study corpus. We calulate the loglikelihood of the word "quarter":

	BNC	Earning.calls	Total.row
Freq of word 'quarter'	7464	27402	34866
Freq of other words	100104111	1889223	101993334
Total column	100104253	1889256	101993509

```r
a <- 7464      # Frequency of 'quarter' in the BNC
b <- 27402     # Frequency 'quarter' in Earning calls corpus
c <- 100104253 # All tokens in BNC
d <- 1889256   # All tokens in Earning calls corpus
```

Expected value corpus 1:

```r
E1 <- c*(a+b)/(c+d)
E1
```

```
## [1] 34220.17
```

Expected value corpus 2:

```r
E2 <- d*(a+b)/(c+d)
E2
```

```
## [1] 645.8333
```

Loglikelihood:

```r
2*((a*log(a/E1)) + (b*log(b/E2)))
```

```
## [1] 182664.9
```

Example 2: the word ‘clean’

	BNC	Earning.calls	Total.row
Freq of word ‘clean’	6512	68	6580
Freq of other words	100097741	1889188	101986929
Total column	100104253	1889256	101993509

```
a <- 6512      # Frequency of 'clean' in the BNC
b <- 68        # Frequency 'clean' in Earning calls corpus
c <- 100104253 # All tokens in BNC
d <- 1889256   # All tokens in Earning calls corpus
```

Expected value corpus 1:

```
E1 <- c*(a+b)/(c+d)
E1
```

```
## [1] 6458.117
```

Expected value corpus 2:

```
E2 <- d*(a+b)/(c+d)
E2
```

```
## [1] 121.8833
```

Loglikelihood:

```
2*((a*log(a/E1)) + (b*log(b/E2)))
```

```
## [1] 28.85126
```

The loglikelihood value of ‘quarter’ is high (182664.9) compared to the one of ‘clean’ (28.85). A large loglikelihood value indicates that a word occures more often in a corpus than could be expected, and hence is a candidate for a keyword in that corpus. In our case, it is not surprising that ‘quarter’ occures more often in a corpus of earning calls transcripts, because quarterly results are a subject of these conversations.