

# word-BERT-rep VS def-BERT-rep (I)

## Objective

Explore similarities between BERT-vector representation (“vector representation” abbreviated as VR) of a word and BERT-vector representation for the word definition. We use average and sum of BERT-VR as baseline for a later comparison with another composition functions.

## Exploring data

We treat with two dataset:

- the first contains (as rows, where name-row is one word) the first 5 words (top 5) with definitions similariest to
- the second have the similarities in before dataset

```
head(data$ranking)
```

```
##           first second  third   fourth    fifth
## love           1  death credit discovery    fuck
## tiger   hundred   seven    five  exhibit   record
## book     hundred   seven    five    record   school
## computer    seven    five viewer  hundred fertility
## plane     hundred   seven    five  exhibit    record
## train     hundred   seven    five  rooster   exhibit
```

```
dim(data$ranking)
```

```
## [1] 202    5
```

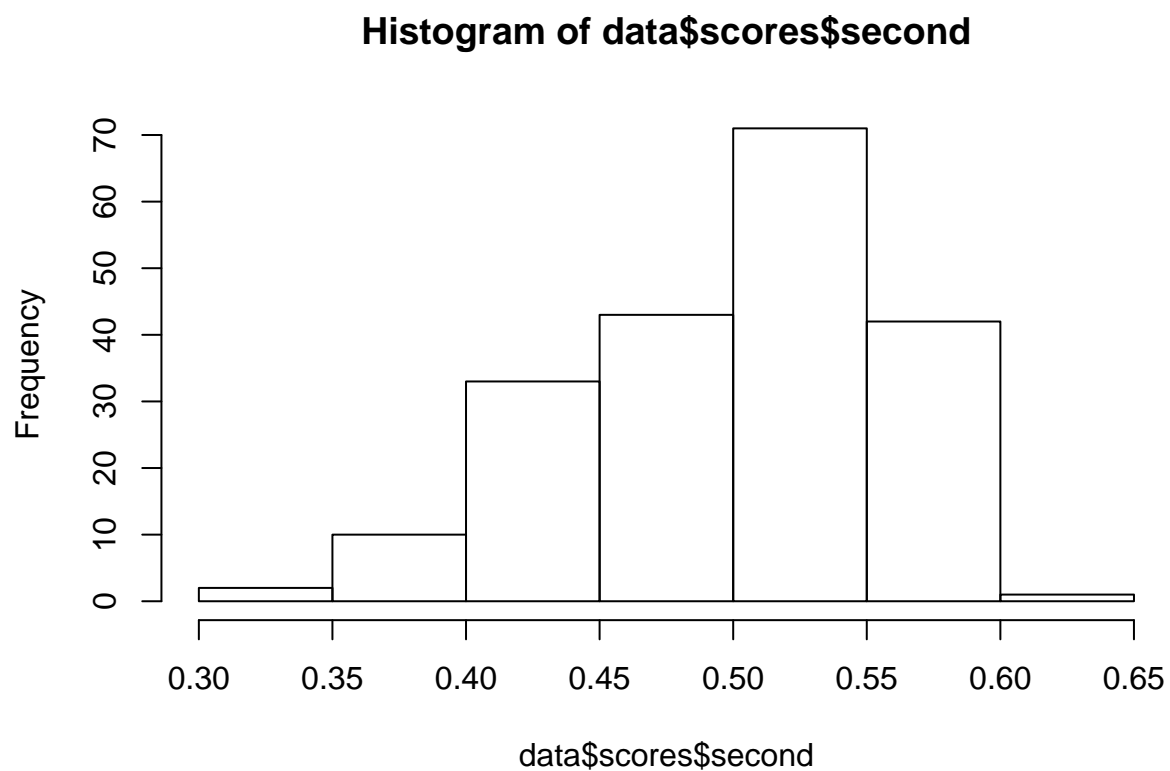
```
head(data$scores)
```

```
##           first    second    third    fourth    fifth
## love    0.4715205 0.4648822 0.4648412 0.4621409 0.4612962
## tiger    0.5680336 0.5422061 0.5059615 0.4415017 0.4293665
## book     0.5449572 0.5376849 0.5355507 0.5160842 0.5083381
## computer 0.5275755 0.5247760 0.5021460 0.4875561 0.4855717
## plane    0.5822837 0.5603127 0.5447720 0.4809039 0.4573219
## train    0.4515287 0.4304824 0.4210125 0.3784637 0.3531368
```

```
dim(data$scores)
```

```
## [1] 202    5
```

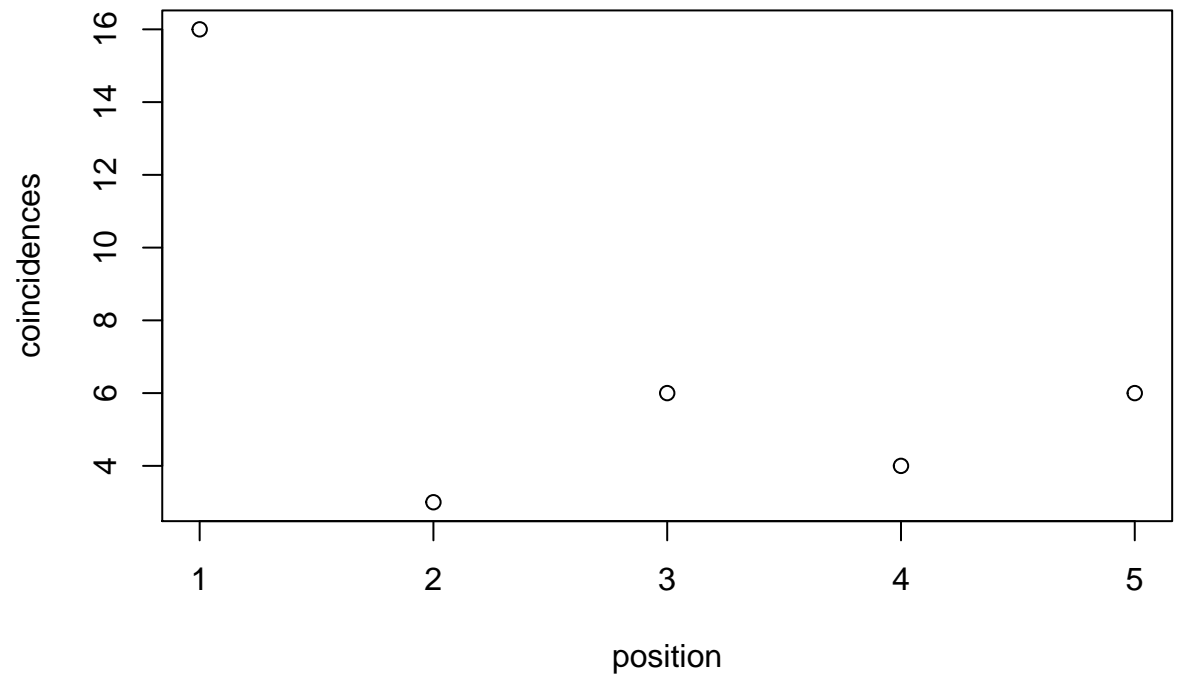
We review the distribution for each position in the ranking



In the five position we can observe the same curve profile (peak in the right). It looks like similarities have a trend to highest similarity values.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that gen-

### coincidences correct word-definition



erated the plot.

Not so much definitions corresponding with word.

We review