test 3: W2V vs BERT - ranking

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
## Warning: package 'stringr' was built under R version 3.4.4
```

```
## Warning: package 'ggplot2' was built under R version 3.4.4
```

DATA

Similarities ranking between words and definitions using average and sum as composition functions for w2v and BERT-free-context.

As example we will show w2v.

```r
# w2v ranking using average as composition function
head(w2v_def_avg_ranking[[1]][ , 1:4])
```

```
##           w        1        2         3
## 1:     love     love   luxury prejudice
## 2:    tiger    tiger     rook      cock
## 3:     book    index  journal   library
## 4: computer software hardware  keyboard
## 5:    plane   flight    plane   airport
## 6:    train    train   voyage    flight
```

```r
# w2v ranking similarities using average as composition function
head(w2v_def_avg_ranking[[2]][ , 1:4])
```

```
##           w            1            2            3
## 1:     love   0.47144239   0.47002609   0.42744012
## 2:    tiger  0.454412687  0.435577771  0.430491449
## 3:     book   0.40064742   0.37588556   0.35384255
## 4: computer   0.68100305   0.53889767   0.49639969
## 5:    plane   0.55917504   0.49426857   0.47720168
## 6:    train  0.358855908  0.346606958  0.343594478
```

```r
# w2v ranking using average as composition function
head(w2v_def_sum_ranking[[1]][ , 1:4])
```

```
##           w        1        2         3
## 1:     love     love   luxury prejudice
## 2:    tiger    tiger     rook      cock
## 3:     book    index  journal   library
## 4: computer software hardware  keyboard
## 5:    plane   flight    plane   airport
## 6:    train    train   voyage    flight
```

```r
# w2v ranking similarities using average as composition function
head(w2v_def_sum_ranking[[2]][ , 1:4])
```

```
##           w            1            2            3
## 1:     love   0.47144239   0.47002609   0.42744012
## 2:    tiger  0.454412687  0.435577771  0.430491449
## 3:     book   0.40064742   0.37588556   0.35384255
## 4: computer   0.68100305   0.53889767   0.49639969
## 5:    plane   0.55917504   0.49426857   0.47720168
## 6:    train  0.358855908  0.346606958  0.343594478
```

Analogous for BERT representation.

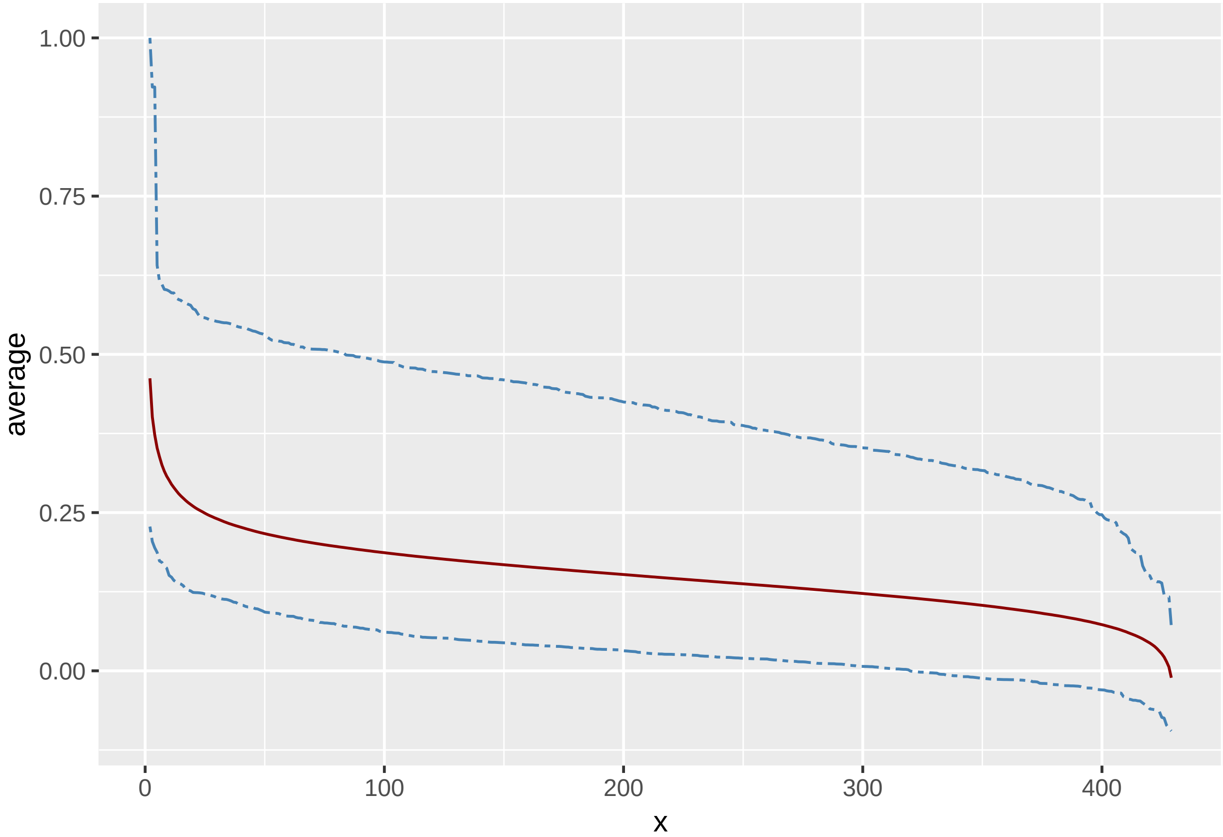
NOTE: the same cosine similarities for average and sum!!!! . . . but really the compound vector is different

RANKING

We study this rankings

w2v compound using words definition average

We can plot the average of similarities with maximum and minimum values,



We can observe that there is similarities equal to one. We review words with simialrity equal to one.

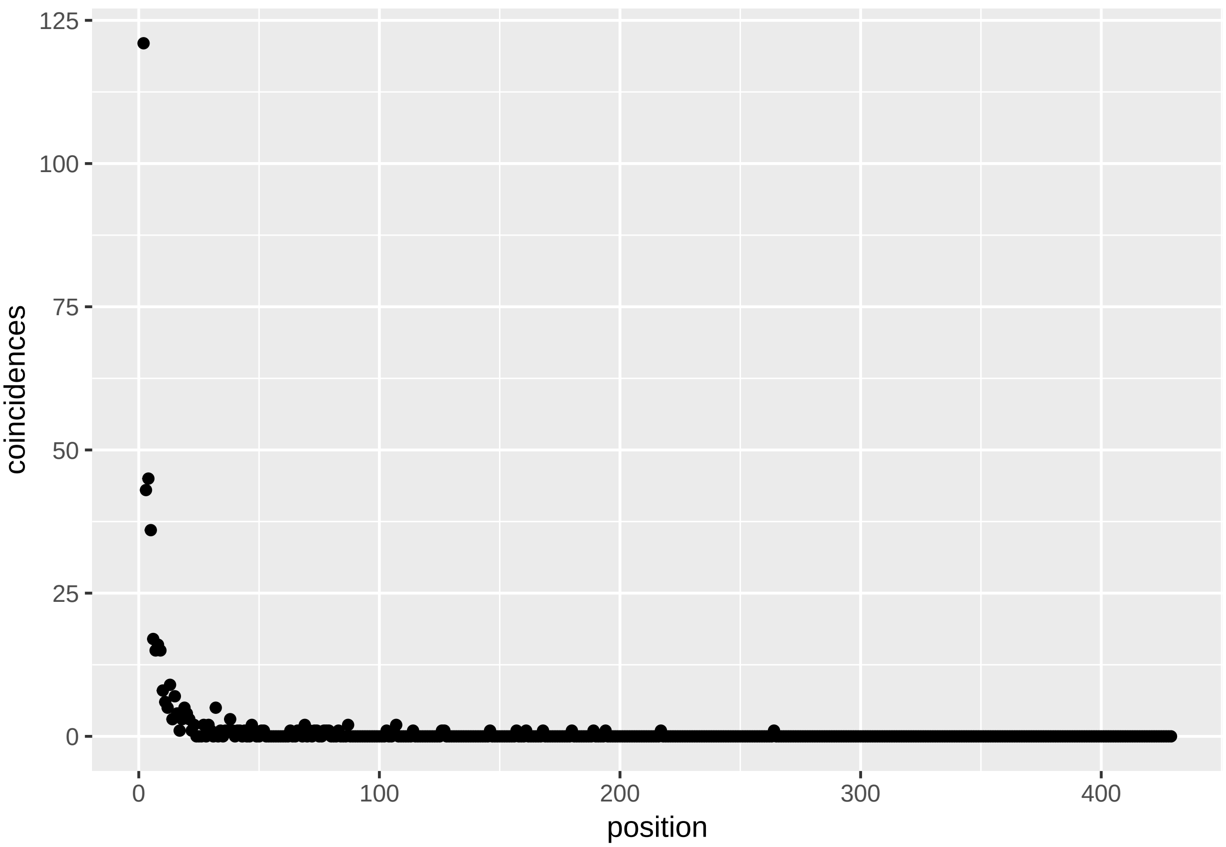
```
sim_scores[as.numeric(sim_scores[["1"]]) == 1, 1:4]
```

```
##              w            1            2            3
## 1:         car    1.00000000    0.55202201    0.45214538
## 2: calculation  1.0000000000  0.4492906353  0.3446186744
## 3:      dollar    1.00000000    0.33452383    0.31103244
```

```
sim_words[w %in% c("car", "calculation", "dollar"), 1:4]
```

```
##              w           1      2           3
## 1:         car  automobile    car       plane
## 2: calculation computation number calculation
## 3:      dollar        buck profit    dividend
```

And we can count the number of definitions equal to word in each ranking position,



125
100
75
50
25
0
coincidences
0
100
200
300
400
position