UNIVERSITÄT LEIPZIG

Alexander von Humboldt
Stiftung/Foundation

Statistical Evaluation II:
Dealing with Context Windows

UNIVERSITÄT LEIPZIG

Alexander von Humboldt
Stiftung/Foundation

Let's start from the beginning

$$LL = \log L(c_{12}, c_1, p) + \log L(c_2\text{-}c_{12}, N\text{-}c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2\text{-}c_{12}, N\text{-}c_1, p_2)$$

c_1 = occurrences of word 1 in the text

c_2 = occurrences of word 2 in the text

c_{12} = co-occurrences of word 1 with word 2 in the text

N = number of tokens in the text

$p = c_2/N$

$p_1 = c_{12}/c_1$

$p_2 = (c_2\text{-}c_{12})/(N\text{-}c_1)$

UNIVERSITÄT LEIPZIG

Alexander von Humboldt
Stiftung/Foundation

Let's start from the beginning

$$LL = \log L(c_{12}, c_1, p) + \log L(c_2\text{-}c_{12}, N\text{-}c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2\text{-}c_{12}, N\text{-}c_1, p_2)$$

c_1 = occurrences of word 1 in the **t e x t**

c_2 = occurrences of word 2 in the **t e x t**

c_{12} = co-occurrences of word 1 with word 2 in the **t e x t**

N = number of tokens in the **t e x t**

$p = c_2/N$

$p_1 = c_{12}/c_1$

$p_2 = (c_2\text{-}c_{12})/(N\text{-}c_1)$

UNIVERSITÄT LEIPZIG

Alexander von Humboldt
Stiftung/Foundation

Let's start from the beginning

$$LL = \log L(c_{12}, c_1, p) + \log L(c_2\text{-}c_{12}, N\text{-}c_1, p) - \log L(c_{12}, c_1, p_1) - \log L(c_2\text{-}c_{12}, N\text{-}c_1, p_2)$$

c_1 = occurrences of word 1 in the **d a t a**

c_2 = occurrences of word 2 in the **d a t a**

c_{12} = co-occurrences of word 1 with word 2 in the **d a t a**

N = number of tokens in the **d a t a**

$p = c_2/N$

$p_1 = c_{12}/c_1$

$p_2 = (c_2\text{-}c_{12})/(N\text{-}c_1)$

Data, not Text!

- We have abstracted data from the text
- We should no longer refer to the text
- But, instead, to the data
- The DataFrames we have constructed have everything we need

Counts for the target word (word 1)

$$f(t) = \frac{1}{W} \sum_{c} n(c, t)$$

t = the target word (word 1)

c = the co-occurrent (word 2)

W = the size of the window

This equation from Bullinaria and Levy, “Extracting Semantic Representations from Word Co-Occurrence Statistics”, 2007,

What does this mean?

$$f(t) = \frac{1}{W} \sum_{c} n(c,t)$$

$n(c,t)$

- Word counts depend on co-occurrence counts!

$$\sum_{c} n(c,t)$$

- Sum all co-occurrence counts for *t*

What does this mean (cont.)?

$$f(t)=\frac{1}{W}\sum_{c} n(c,t)$$

$$\frac{1}{W}\sum_{c} n(c,t)$$

- Finally, divide by the total window size (*L* + *R*)

Counts for the co-occurrent and *N*

$$f(c) = \frac{1}{W} \sum_t n(c, t)$$

- Sum of the co-occurrences of *c* with every *t*

$$N = \frac{1}{W} \sum_t \sum_c n(c, t)$$

- Sum of the counts for every *t*