Correlation

Prof. Eric A. Suess

April 5, 2021

Counting and correlating pairs of words

- Tokenizing by n-gram is a useful way to explore pairs of adjacent words.
- Tidy data is a useful structure for comparing between variables or grouping by rows, but it can be challenging to compare between rows: for example, to count the number of times that two words appear within the same document, or to see how correlated they are.
- Most operations for finding pairwise counts or correlations need to turn the data into a **wide matrix** first.

Wide Format to examine correlation

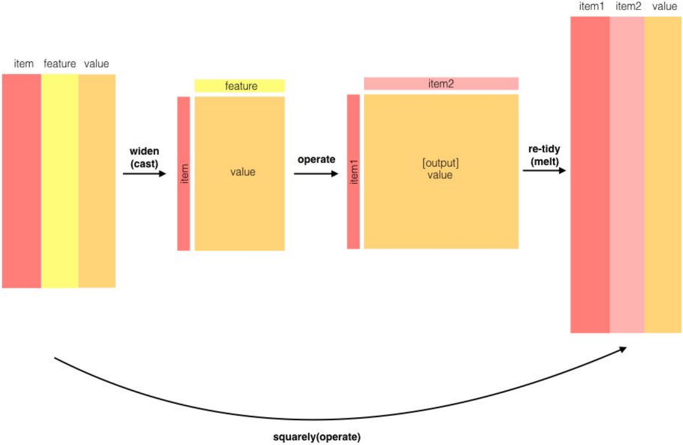

Figure 1:

Phi Coeffient

- "We may instead want to examine correlation among words, which indicates how often they appear together relative to how often they appear separately."
- See the Wikipedia page about the Phi Coefficient