Notes about correlation (for Asgn 2)

Sharon Goldwater

School of informatics

Overview of assignment

Exploration of distributional similarity.

- Work with data extracted from Twitter (co-occurrence counts)
- Compare different ways to contruct context vectors and compute similarities
- Analyze and discuss differences between approaches, qualitatively and quantitatively.

Work through the lab **before** you start the assignment!

Qualitative and quantitative analysis

Assignment asks you to do some of each.

- Examples of qualitative analysis:
  - Using visualization to illustrate/discuss examples or trends
  - Discussing one or a few examples in more detail, by looking at our dataset and/or other Tweets (e.g., use the Twitter search page).

- Examples of quantitative analysis:
  - Often: numerical comparison to a gold standard of accuracy
  - Here: consider other options, such as correlating similarity measures against word frequency.

One kind of quantitative analysis

- Assignment spec suggests you may want to consider **correlation** between similarity measures and word frequency.

- Why?
  - A good similarity measure should measure (only) similarity.
  - So presumably *not* be correlated with frequency.
  - Unless more frequent words really are more similar to each other! (Would need to test with humans... let's assume not)

What is correlation?

- Intuitively: two random variables X and Y are **correlated** if, when the value of X increases, the value of Y also tends to increase (positive correlation) or decrease (negative correlation).

- Often, X and Y are different measurements for each data point.
  - A person's height X and weight Y
  - A word's frequency X and length Y

- Two standard ways to measure correlation:
  - Spearman (rank) correlation: roughly as above.
  - Pearson (linear) correlation: more specific.

Pearson correlation

- Mathematically: the covariance of X and Y, normalized by the product of their individual standard deviations.

- Intuitively: if I plot X against Y, how close to a perfect linear relationship do I see?
  - Does not measure the *slope* of the line, just whether there is one. (Compare rows 1 and 2, next page.)
  - Does not tell us if there's some other *non-linear* relationship between X and Y. (See row 3, next page.)

- For data samples, the Pearson correlation coefficient is usually denoted r.

Pearson correlation

Examples datasets with Pearson r values shown:

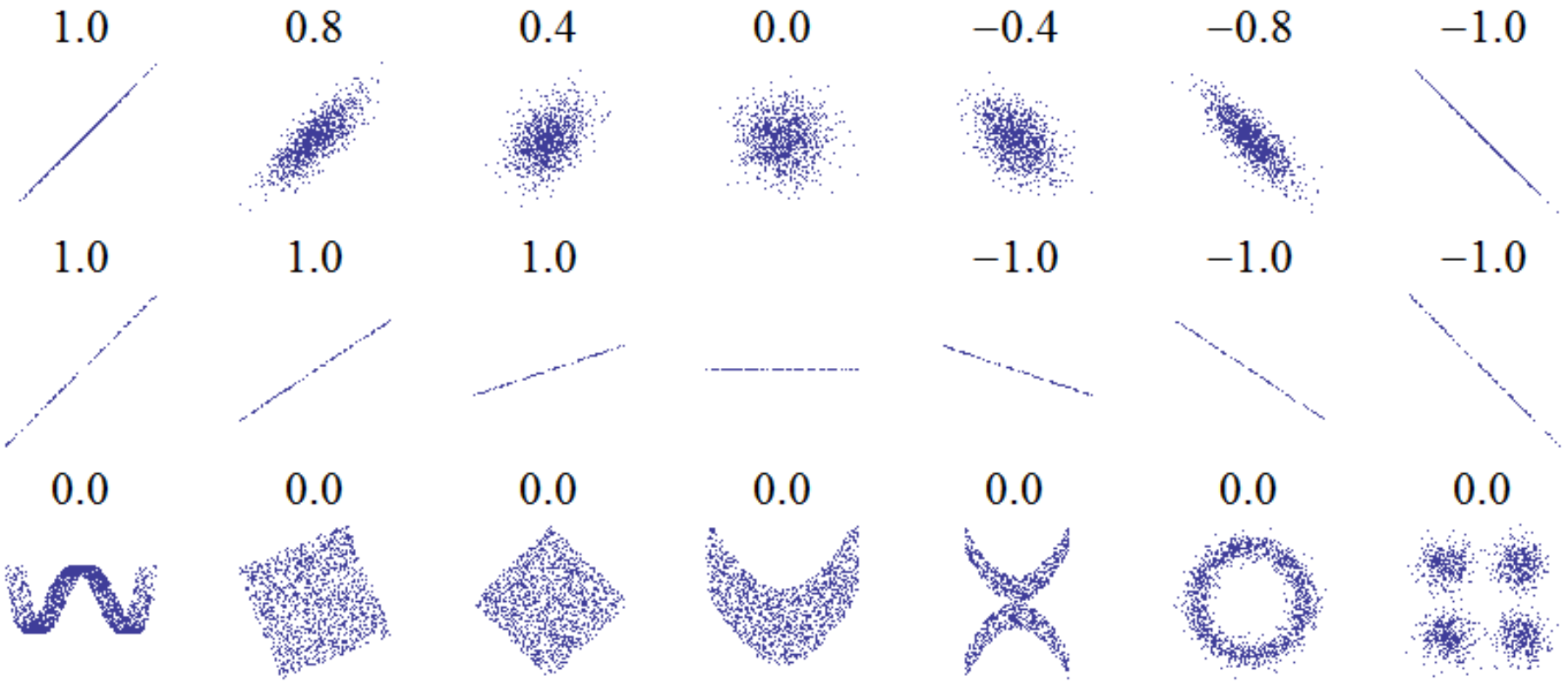


Image source: `https://commons.wikimedia.org/wiki/File:Correlation_examples.png`

Spearman rank correlation

- Mathematically: compute the Pearson correlation between the **rank ordering** of X and Y values.
- Intuitively: how close to a perfectly monotonic relationship do X and Y have? (i.e., when X increases, Y increases)
- For data samples, the Spearman rank correlation coefficient is usually denoted ρ or r_s.

Spearman correlation

Data with perfect rank correlation, but not perfectly linear:

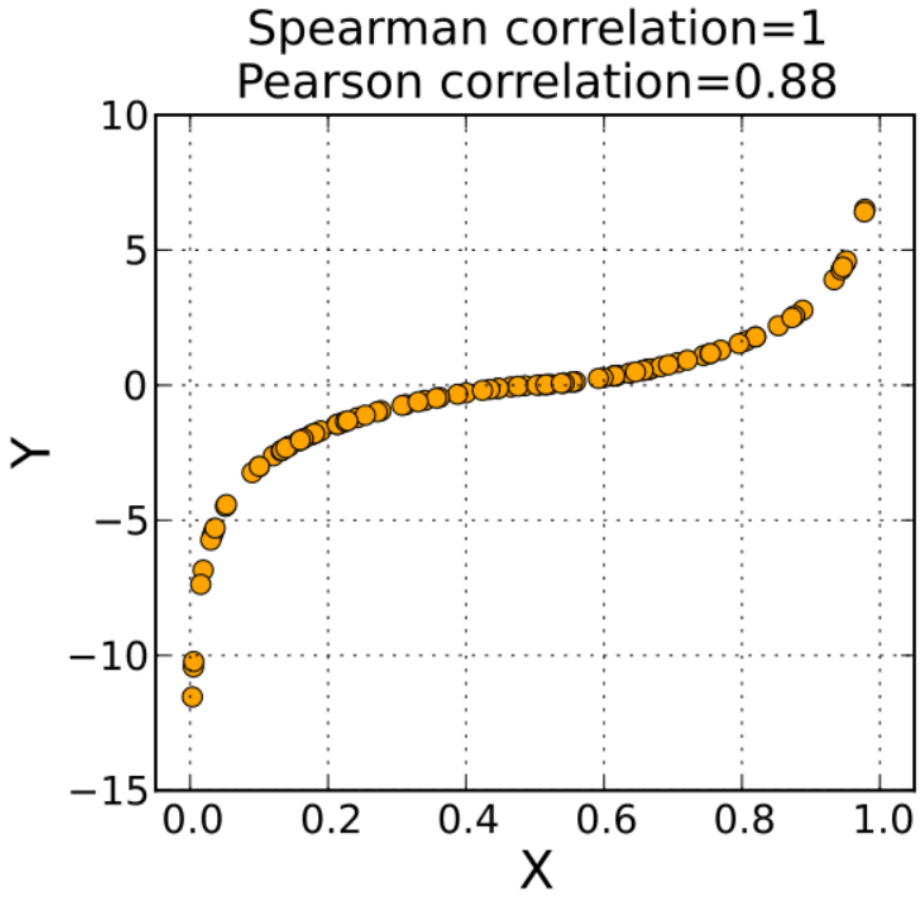


Image by Skbkekas (CC-BY-SA 3.0)

`https://en.wikipedia.org/wiki/Spearman\%27s_rank_correlation_coefficient`

Which one to use?

- If correlation is roughly linear, Pearson will normally yield stronger results (larger absolute values)
  - If hypothesis testing against the possibility of no correlation, likely to have higher significance level than Spearman.
  - But if using large samples from corpora, often nearly *any* result is clearly "non-zero". We may care more about the actual degree of correlation.
- If correlation is non-linear, or nothing is known, use Spearman.

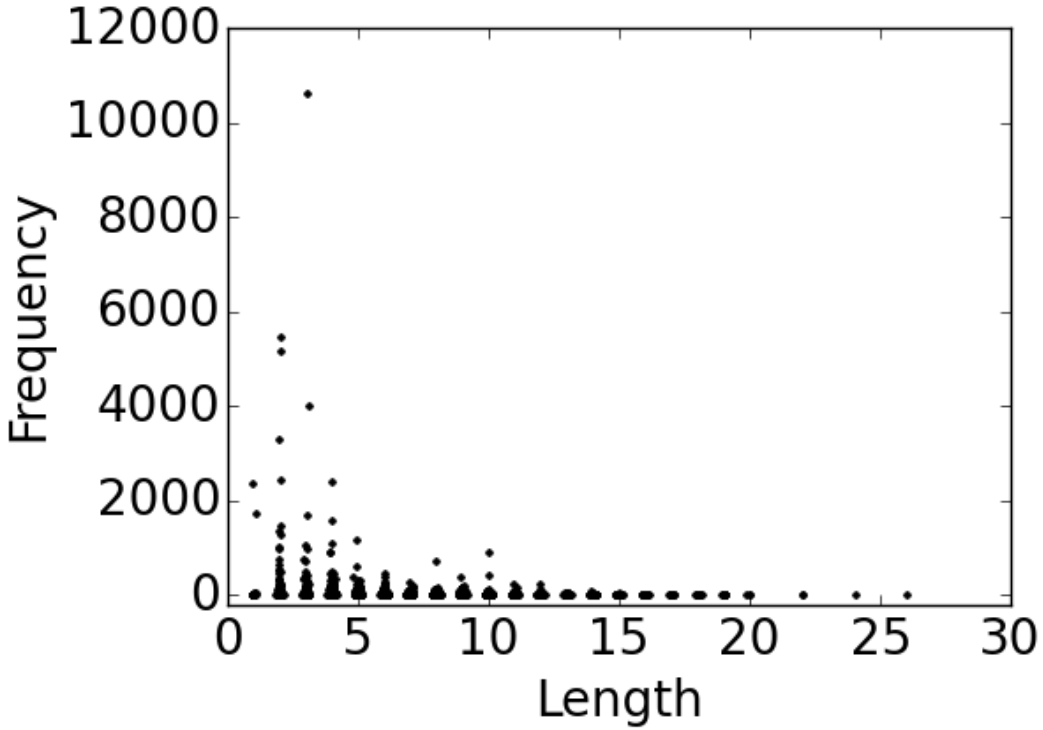
But usually we do know something

Best to look at the data first! For example, word freq vs length:

Seems to follow a pattern, but not strongly linear. Indeed,

- Spearman: $\rho = -0.18$
- Pearson: $r = -0.10$

(Note: I "jittered" the data so those with same (x,y) are not right on top of each other.)



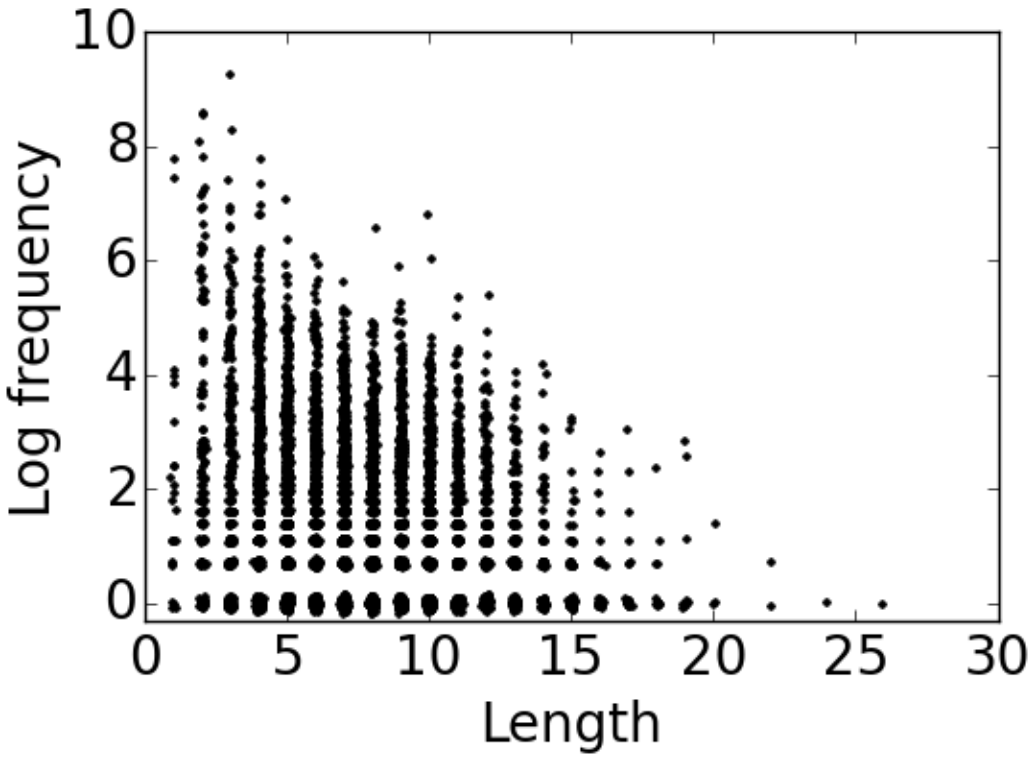
Log frequency

Of course, using log frequencies is often more sensible:

We now have

- Spearman: $\rho = -0.18$
- Pearson: $r = -0.21$

Notice that ρ is not affected by rescaling the data. r is higher, but still only a weak linear correlation.



So, which one to use?

- So, Pearson can still work if there is an obvious tranformation to make the correlation roughly linear.
- But if in doubt, usually fine to use Spearman.
- As with all statistics, many subtleties if using for really careful analysis (see statistics course or online tutorials), but what I've said is probably enough for exploratory studies (i.e., your assignment).