Estimating Word Similarity: Instrinsic Evaluation

Intrinsic *Evaluation* of Similarity/Relatedness Estimations

- *Spearman's rank Correlation Coefficient* (Spearman's ρ)

$$\rho = 1 - \frac{6 \sum_{i_1}^{n} (u_i - v_i)^2}{n(n^2 - 1)}$$

[?]

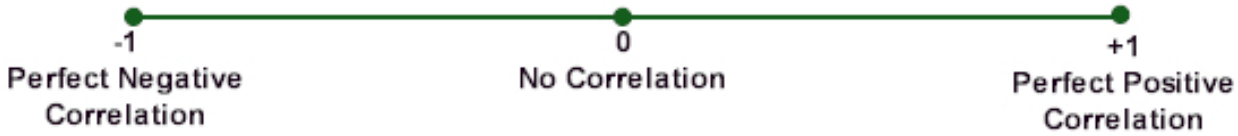
- Compares ranking of n items, u_i and v_i. I.e., the measure assesses monotonic relationships
- Here: Compare ranking of word pairs that is based on sim/rel values:
  - for word pairs $p_1, \ldots p_n$,
  - sim/rel judgements elicited from humans $(u_1, \ldots u_n)$
  - vs. algorithm's estimated similarity scores $(v_1, \ldots v_n)$

$\Rightarrow$ How different is the estimated ranking to the human-based ranking?

Estimating Word Similarity: Instrinsic Evaluation

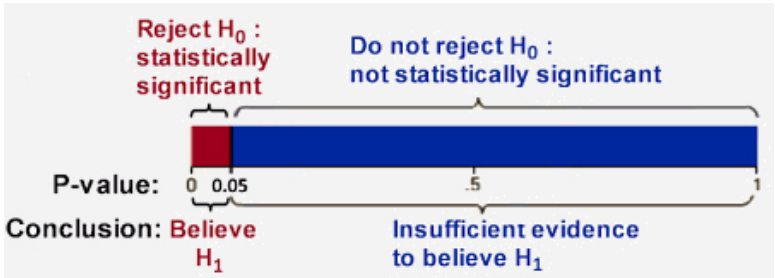
Spearman's rank Correlation Coefficient (Spearman's ρ)

- Can model simulate human's ability to judge word similarity?
- Null Hypothesis (H0): There is no correlation between estimated and human-based judgements
- H1: There is a correlation, i.e., model is a weak/moderate/strong/very strong estimator
- Interpretation of Spearman's ρ value



Figures: `https://geographyfieldwork.com/SpearmansRankCalculator.html`

- p-value (probability value): Measure of how likely or probable it is that any observed correlation is due to chance.



Range between 0 (0%) and 1 (100%)

Estimating Word Similarity: Instrinsic Evaluation

Example 1

Spearman's $\rho = 0.257$ (p-value$= 0.37$)

No.	Word Pair	Predicted	Reference
0	easy-difficult	0.83	0.58
1	simple-easy	0.79	9.40
2	bad-great	0.64	0.35
3	difficult-simple	0.69	0.87
4	bad-terrific	0.61	0.65
5	dinner-breakfast	0.89	3.33
6	meal-dinner	0.74	7.15
7	boat-car	0.60	2.37
8	sandwich-lunch	0.58	6.30
9	heroine-hero	0.69	8.78
10	car-gauge	0.46	1.13
11	wagon-carriage	0.74	7.70
12	car-carriage	0.65	5.13
13	meal-waist	0.14	0.98

$\Rightarrow$ We cannot reject the Null hypothesis.

Estimating Word Similarity: Instrinsic Evaluation

Example 2

Spearman's $\rho = 0.65$ (p-value$= 0.04$)

No.	Word Pair	Predicted	Reference
	~~easy-difficult~~	~~0.83~~	~~0.58~~
0	simple-easy	0.79	9.40
	~~bad-great~~	~~0.64~~	~~0.35~~
	~~difficult-simple~~	~~0.69~~	~~0.87~~
	~~bad-terrific~~	~~0.61~~	~~0.65~~
1	dinner-breakfast	0.89	3.33
2	meal-dinner	0.74	7.15
3	boat-car	0.60	2.37
4	sandwich-lunch	0.58	6.30
5	heroine-hero	0.69	8.78
6	car-gauge	0.46	1.13
7	wagon-carriage	0.74	7.70
8	car-carriage	0.65	5.13
9	meal-waist	0.14	0.98

$\Rightarrow$ We can reject the Null hypothesis.