



di.unito.it

DIPARTIMENTO  
DI INFORMATICA

TLN-LAB

utilizzo di risorse  
lessicografiche per la concept  
similarity e la WSD

↑  
Word Sense  
Disambiguation

Daniele Radicioni

# credits

- the following slides have been mostly built on materials from:
  - M. Lesk. Automatic Sense Disambiguation using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th International Conference on Systems Documentation*, 1986.
  - Tanveer Siddiqui and U.S.Tiwary, *Natural Language Processing and Information Retrieval*, Oxford University, 2008.

# conceptual similarity with WordNet



Daniele Radicioni - TLN

# conceptual similarity with WN

- dati in input due termini, il task di conceptual similarity consiste nel fornire un punteggio numerico di similarità che ne indichi la vicinanza semantica.
  - ad esempio, la similarità fra i concetti car e bus potrebbe essere 0.8 in una scala [0,1], in cui 0 significa che i sensi sono completamente dissimili, mentre 1 significa identità.
- per risolvere il task di conceptual similarity è possibile sfruttare la struttura ad albero di WordNet.

# input

- l'input per questa esercitazione è costituito da coppie di termini contenute nel file WordSim353 ([disponibile nei formati .tsv e .csv](#))
  - Il file contiene 353 coppie di termini utilizzati come testset in varie competizioni internazionali
  - A ciascuna coppia è attribuito un valore numerico [0, 10], che rappresenta la similarità fra gli elementi della coppia.

# consegna

- l'esercitazione consiste nell'implementare tre misure di similarità basate su WordNet.
- per ciascuna di tali misure di similarità, calcolare gli indici di correlazione di Spearman and gli indici di correlazione di Pearson fra i risultati ottenuti e quelli 'target' presenti nel file annotato.



**WIKIPEDIA**  
The Free Encyclopedia

Article

Talk

# Pearson correlation coefficient

From Wikipedia, the free encyclopedia

## Definition [ edit ]

Pearson's correlation coefficient is the [covariance](#) of the two variables divided by the product of their [standard deviations](#). The form of the definition involves a "product moment", that is, the mean (the first [moment](#) about the origin) of the product of the mean-adjusted random variables; hence the modifier *product-moment* in the name.

## For a population [ edit ]

Pearson's correlation coefficient when applied to a [population](#) is commonly represented by the Greek letter  $\rho$  (rho) and may be referred to as the *population correlation coefficient* or the *population Pearson correlation coefficient*. Given a pair of random variables ( $X, Y$ ), the formula for  $\rho$ <sup>[7]</sup> is:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \quad (\text{Eq.1})$$

← non entriamo nel merito, le prendiamo  
= come sono

where:

- cov is the [covariance](#)
- $\sigma_X$  is the [standard deviation](#) of  $X$
- $\sigma_Y$  is the standard deviation of  $Y$





WIKIPEDIA  
The Free Encyclopedia

Article Talk

# Spearman's rank correlation coefficient

From Wikipedia, the free encyclopedia

## Definition and calculation [ edit ]

The Spearman correlation coefficient is defined as the [Pearson correlation coefficient](#) between the [rank variables](#).<sup>[3]</sup>

For a sample of size  $n$ , the  $n$  [raw scores](#)  $X_i, Y_i$  are converted to ranks  $\text{rg } X_i, \text{rg } Y_i$ , and  $r_s$  is computed from:

$$r_s = \rho_{\text{rg}_X, \text{rg}_Y} = \frac{\text{cov}(\text{rg}_X, \text{rg}_Y)}{\sigma_{\text{rg}_X} \sigma_{\text{rg}_Y}}$$

where

- $\rho$  denotes the usual [Pearson correlation coefficient](#), but applied to the rank variables.
- $\text{cov}(\text{rg}_X, \text{rg}_Y)$  is the [covariance](#) of the rank variables.
- $\sigma_{\text{rg}_X}$  and  $\sigma_{\text{rg}_Y}$  are the [standard deviations](#) of the rank variables.

# ① Wu & Palmer *sono sensi, non termini* → si usa WordNet

$$cs(s_1, s_2) = \frac{2 \cdot depth(\text{LCS})}{depth(s_1) + depth(s_2)}$$

- la misura di similarity di Wu & Palmer si basa sulla struttura di WordNet
- LCS è il primo antenato comune (Lowest Common Subsumer) fra i sensi  $s_1$  e  $s_2$ ; e depth(x) è una funzione che misura la distanza fra la radice di WordNet e il synset  $x$ .

## ② Shortest Path

$$\text{sim}_{\text{path}}(s_1, s_2) = 2 \cdot \text{depthMax} - \text{len}(s_1, s_2)$$

- for a specific version of WordNet,  $\text{depthMax}$  is a fixed value.
- the similarity between two senses  $(s_1, s_2)$  is the function of the shortest path  $\text{len}(s_1, s_2)$  from  $s_1$  to  $s_2$ .
- if  $\text{len}(s_1, s_2)$  is 0,  $\text{sim}_{\text{path}}(s_1, s_2)$  gets the maximum value of  $2 * \text{depthMax}$ .
- if  $\text{len}(s_1, s_2)$  is  $2 * \text{depthMax}$ ,  $\text{sim}_{\text{path}}(s_1, s_2)$  gets the minimum value of 0.
- thus, the values of  $\text{sim}_{\text{path}}(s_1, s_2)$  are between 0 and  $2 * \text{depthMax}$ .

### ③ Leakcock & Chodorow


$$\text{sim}_{LC}(s_1, s_2) = - \log \frac{\text{len}(s_1, s_2)}{2 \cdot \text{depthMax}}$$

- when  $s_1$  and  $s_2$  have the same sense,  $\text{len}(s_1, s_2) = 0$ . in practice, we add  $l$  to both  $\text{len}(s_1, s_2)$  and  $2 * \text{depthMax}$  to avoid  $\log(0)$ .
- thus the values of  $\text{sim}_{LC}(s_1, s_2)$  are in the interval  $(0, \log(2 * \text{depthMax} + l)]$

# termini vs. sensi

- attenzione: l'input è costituito da coppie di termini, mentre la formula utilizza sensi.
- per calcolare la similarity fra 2 termini immaginiamo di prendere la massima similarity fra tutti i sensi del primo termine e tutti i sensi del secondo termine.
- l'ipotesi è cioè che i due termini funzionino come contesto di disambiguazione l'uno per l'altro.
- nella formula c sono i concetti che appartengono ai synset associati ai termini  $w_1$  e  $w_2$ .

. calcio - potash  
. calcio - Ronaldinho  
. calcio - pistola  
. calcio - lotte

$$\text{sim}(w_1, w_2) = \max_{c_1 \in s(w_1), c_2 \in s(w_2)} [\text{sim}(c_1, c_2)]$$

*Daniele Radicioni - ILN*