

# Taxonomy-based feature dissimilarity measure

## Core Idea:

Compute semantic distance using all taxonomic ancestors of concepts in an ontology, not just the shortest path.

## Method:

1. For a concept  $a$ , collect all its taxonomic ancestors  $\rightarrow$  set  $\phi(a)$
2. For concept  $b$ , collect  $\phi(b)$
3. Compare sets to count:
  - **Unique features** (in one but not the other)
  - **Common features** (in both)
4. Calculate **dissimilarity**:

$$dis(a, b) = \log_2 \left( 1 + \frac{|unique_a| + |unique_b|}{|unique_a| + |unique_b| + |common|} \right)$$

- Result is in  $[0, 1]$ , where 0 means identical and 1 means maximally different.
- **Handles Polysemy:**

For words with multiple senses, dissimilarity is computed as:

$$dissim(W_1, W_2) = \min_{i,j} dissim(a_i, b_j)$$

where  $a_i$  and  $b_j$  are the senses of  $W_1$  and  $W_2$  respectively.

## Sentence-Level Dissimilarity

$$D(A, B) = \frac{1}{|A|} \sum_{e_a \in A} \min_{e_b \in B} dis(e_a, e_b)$$

- For each entity in sentence A, pick the most similar entity in B.
- Average over all entities in A.