ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nghĩa từ vựng và phân giải nhập nhằng từ

Viện CNTT &TT – Trường ĐHBKHN

Từ đồng âm

- Từ đồng âm (Homonymy): là những từ trùng nhau về hình thức ngữ **âm** nhưng khác nhau về **nghĩa**
  - Từ đồng âm, đồng tự (Homograph) : các từ với cùng cách viết nhưng có nghĩa khác nhau. Ví dụ:
    - dove - dive into water, white bird
    - saw
  - Từ đồng âm, không đồng tự (Homophone): các từ có cách viết khác nhau nhưng có cùng âm. Ví dụ:
    - see, sea; meat, meet

Phân loại từ đồng âm tiếng Việt

- Đồng âm từ với từ, gồm:
  - Đồng âm từ vựng: Tất cả các từ đều thuộc cùng một từ loại. Ví dụ:
    - *đường*$_1$ (đắp đường) - *đường*$_2$ (đường phèn).
    - *đường kính*$_1$ (đường để ăn) - *đường kính*$_2$ (…của đường tròn).
    - *cất*$_1$ (cất vó) - *cất*$_2$ (cất tiền vào tủ) - *cất*$_3$ (cất hàng) - *cất*$_4$ (cất rượu)
  - Đồng âm từ vựng-ngữ pháp: Các từ trong nhóm đồng âm với nhau chỉ khác nhau về từ loại. Ví dụ:
    - *chỉ*$_1$ (cuộn chỉ) - *chỉ*$_2$ (chỉ tay năm ngón) - *chỉ*$_3$ (chỉ còn có dăm đồng).
    - *câu*$_1$ (nói vài câu) - *câu*$_2$ (rau câu) - *câu*$_3$ (chim câu) - *câu*$_4$ (câu cá)
- Đồng âm từ với tiếng: các đơn vị khác nhau về cấp độ; kích thước ngữ âm của chúng đều không vượt quá một tiếng. Ví dụ:
  - Con trai Văn *Cốc* lên dốc bắn cò, đứng lăm *le* cười khanh *khách*. Con gái Bát *Tràng* bán hàng thịt ếch ngồi châu *chẫu* nói ương *ương*.

Từ đa nghĩa, đồng nghĩa

- Từ đa nghĩa (Polysemy): một từ có thể có nhiều nghĩa mà cú pháp chỉ giúp phân biệt nghĩa đ/v các từ loại khác nhau của 1 từ nhập nhằng
  - *chỉ*$_1$ (cuộn chỉ) - *chỉ*$_2$ (chỉ tay năm ngón) - *chỉ*$_3$ (chỉ còn có dăm đồng).
  - "conduct" (noun or verb)
    - John's conduct in class is unacceptable.
    - John will conduct the orchestra on Thursday.
- Đồng nghĩa (Synonymy): là những từ tương đồng với nhau về nghĩa, khác nhau về âm thanh. Ví dụ
  - cố, gắng
  - car, automobile

Nghĩa từ vựng

- Nghĩa của 1 từ là gì?
  - Homonyms (các nghĩa khác nhau)
    - bank: financial institution
    - bank: sloping land next to a river
  - Polysemes (các nghĩa có liên quan/gần nhau)
    - bank: financial institution as corporation
    - bank: a building housing such an institution
- Các nguồn ngữ liệu đ/v nghĩa từ:
  - Dictionaries (thesaurus)
  - Lexical databases

Nghĩa từ vựng

- Ngữ nghĩa nghiên cứu ý nghĩa của các phát biểu dạng ngôn ngữ
- Nghĩa từ vựng (Lexical semantics) nghiên cứu:
  - quan hệ từ vựng: sự liên hệ về mặt ngữ nghĩa giữa các từ
  - ràng buộc về lựa chọn: cấu trúc liên hệ ngữ nghĩa bên trong của từng từ
  - bao gồm lý thuyết về:
    - phân loại và phân rã nghĩa của từ
    - sự giống và khác trong cấu trúc từ vựng – ngữ nghĩa giữa các ngôn ngữ
    - quan hệ nghĩa của từ với cú pháp và ngữ nghĩa của câu.

Các ứng dụng

- Dịch máy
- Tóm tắt văn bản
- Phân loại văn bản
- Phân tích quan điểm
- Quảng cáo hướng ngữ cảnh
- Đối sánh văn bản
- Máy tìm kiếm
- Hệ thống hội thoại (dialogue system)
- Hệ thống hỏi đáp (question answering)

Ràng buộc về lựa chọn

- Có rất nhiều từ đòi hỏi các bổ nghĩa (thường là các Động từ- các vị từ). Các bổ nghĩa này thường là các Danh từ và phải thỏa mãn các ràng buộc về lựa chọn.
- Ví dụ:
  - read (human subject, textual object)
  - eat (animate subject)
  - kill (animate object)
- Sử dụng vị từ để phân giải nhập nhằng ?
  - Một kiểu thông tin ngữ cảnh là thông tin về kiểu các bổ nghĩa mà 1 từ nhập nhằng yêu cầu.
  - Các vị từ khác nhau ứng với các nghĩa khác nhau
    - wash the **dishes** (theme : washable-thing)
    - serve vegetarian **dishe**s (theme : food-type)
  - Kiểu các bổ nghĩa cũng có thể giải quyết nhập nhằng cho vị từ

Đánh giá về các ràng buộc

- Yêu cầu liệt kê đầy đủ trong dạng máy có thể đọc được:
  - Cấu trúc bổ nghĩa của các Động từ.
  - Các ràng buộc về lựa chọn của các bổ nghĩa.
  - Mô tả các đặc tính của các từ đáp ứng được tiêu chí của ràng buộc về lựa chọn.
    - E.g. This flight serves the "**region"** between Mumbai and Delhi
    - How do you decide if "region" is compatible with "sector"
  - Sử dụng Từ điển đồng nghĩa hay Wordnet:
    - gồm từ đồng nghĩa (Synonyms) và trái nghĩa (Antonyms)
    - Từ lớp cha và từ lớp con
- Độ chính xác:
  - 44% on Brown corpus.

Đánh giá về các ràng buộc

- Các danh từ riêng (tên riêng) trong ngữ cảnh của 1 từ nhập nhằng có thể xem như dấu hiện xử lý nhập nhằng rất mạnh.

  E.g. **"Sachin Tendulkar"** will be a strong indicator of the category **"sports".**

  Sachin Tendulkar plays **cricket.**

- Các danh từ riêng không xuất hiện trong thesaurus hay Wordnet. Từ đó cách tiếp cận này không khai thác được các dấu hiệu mạnh của các danh từ riêng.
- Độ chính xác:
  - 50% khi được test trên 10 từ có nhiều nghĩa trong tiếng Anh.

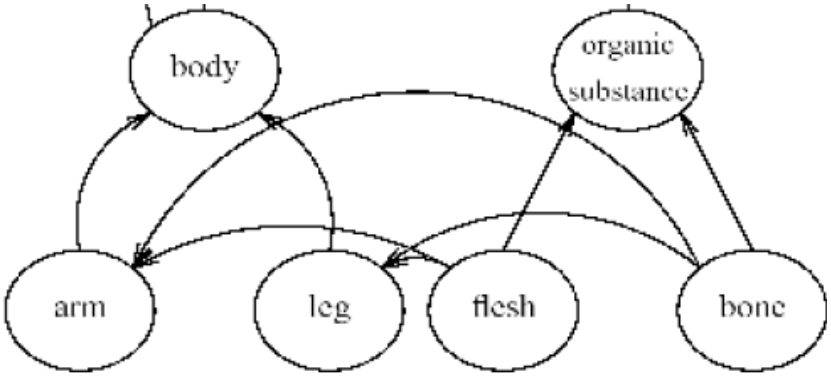
Đánh giá về các ràng buộc

- Ưu điểm
  - Một tiếp cận không phân tích cú pháp.
  - Cài đặt đơn giản.
  - Không yêu cầu 1 bộ dữ liệu đ/v từ nhập nhằng.
- Nhược điểm
  - Có thể gặp đối sánh thưa*:* khả năng bao trùm từ là rất ít.
  - Không sử dụng được với các trường hợp không liệt kê trong máy.
  - Các danh từ riêng (tên riêng) trong ngữ cảnh của 1 từ nhập nhằng có thể xem như dấu hiện xử lý nhập nhằng rất mạnh nhưng các danh từ riêng không xuất hiện trong thesaurus. Từ đó cách tiếp cận này không khai thác được các dấu hiệu mạnh của các danh từ riêng.

Đánh giá về các ràng buộc

- Vấn đề:
  - Đôi khi ràng buộc lựa chọn không đủ chặt (khi 1 từ có nhiều nghĩa)
  - Đôi khi ràng buộc quá chặt – khi vị từ sử dụng phép ẩn dụ. Vd, I'll eat my hat!

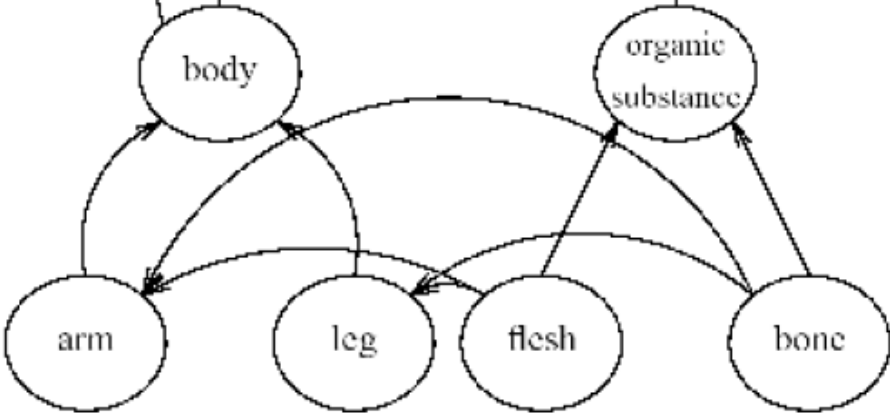
WordNet: Giới thiệu



CSDL từ vựng

- Xây dựng một mạng khổng lồ các từ vựng và quan hệ giữa các từ vựng
- Wordnet tiếng Anh
  - 4 lớp: danh từ, động từ, tính từ, trạng từ
  - Danh từ: 120,000; Động từ: 22,000; Tính từ: 30,000;
  - Trạng từ: 6,000

WordNet: Giới thiệu



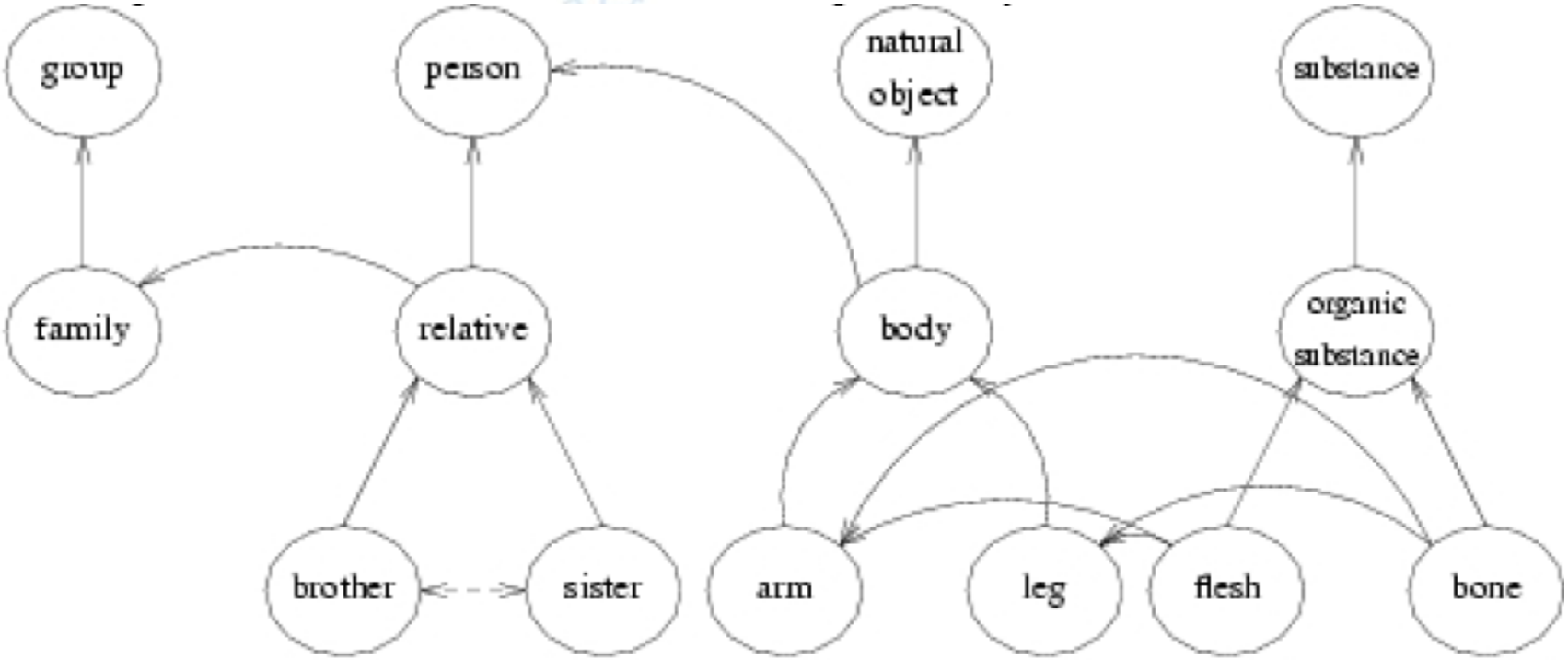
- CSDL từ vựng
  - Wordnet cho các ngôn ngữ khác [www.globalwordnet.org]
    - Có wordnet cho các ngôn ngữ: Tây Ban Nha, Tiệp, Hà Lan, Pháp, Đức, Ý, Bồ Đào Nha, Thụy Điển, Basque, Estonian
    - Wordnets đang được làm cho các tiếng: Bulgary, Đan mạch, Hy lạp, Hebrew, Hindi, Cannada, Latvian, Moldavy, Romany, Nga, Slovenian, Tamil, Thái lan, Thổ Nhĩ Kỳ, Ireland, Nauy, Ba tư, Iran

Tập từ đồng nghĩa
Synonym Sets - Synsets

- Từ có nhập nhằng
- Các nút trong Wordnet biểu diễn tập từ đồng nghĩa “synonym sets”, hoặc *synsets.* Ví dụ:
  - Fool: 1 người dễ bị lợi dụng
  - {chump, fish, fool, gull, mark, patsy, fall guy, sucker, schlemiel, shlemiel, soft touch, mug}
  - Synset = tập khái niệm

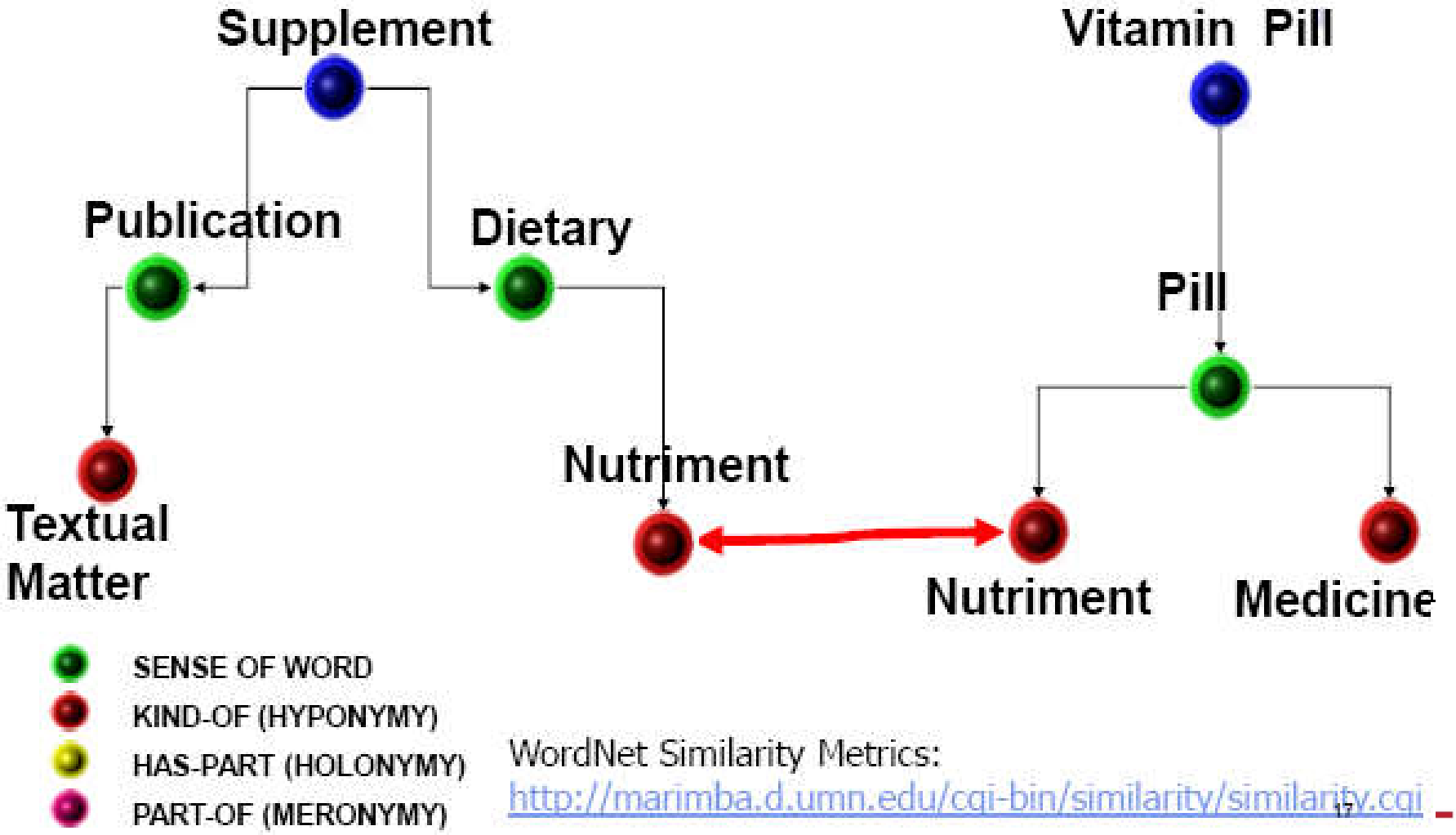
Các quan hệ khác trong WordNet

- Các từ nối theo chiều dọc biểu diễn quan hệ rộng (holonymy) - hẹp (hypernymy), theo chiều ngang biểu diễn quan hệ bộ phận meronymy (part_of) và holonymy (has_part) .
- Mỗi nghĩa của từ được biểu diễn bằng 1 số synset



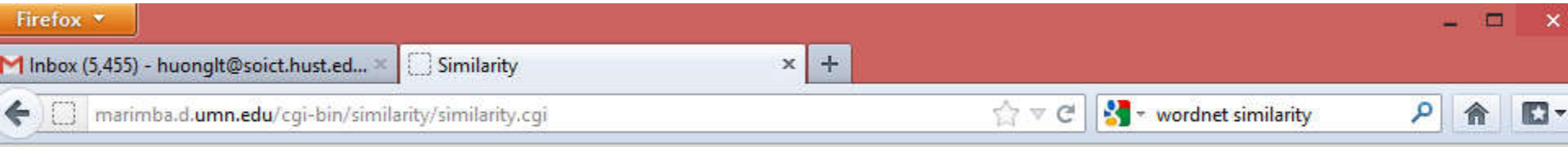
http://wordnet.princeton.edu/

Phân giải nhập nhằng sử dụng quan hệ từ vựng



WordNet Similarity Metrics:
http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi

17



WordNet::Similarity

Read an overview of WordNet::Similarity.

You may enter any two words in one of three formats:

1. `word`
2. `word#part_of_speech` (where part_of_speech is one of n, v, a, or r)
3. `word#part_of_speech#sense` (where sense is a positive integer)

If words are entered in format 1 or 2, then the relatedness of all valid forms of the words will be computed (e.g., if 'dogs' is entered, then 'dog' will be used to compute relatedness). More instructions.

Word 1: ( ) Use all senses ( ) Pick a sense by gloss ( ) Pick a sense by synset

Word 2: ( ) Use all senses ( ) Pick a sense by gloss ( ) Pick a sense by synset

Measure: Path Length About the measures

[x] Use root node?

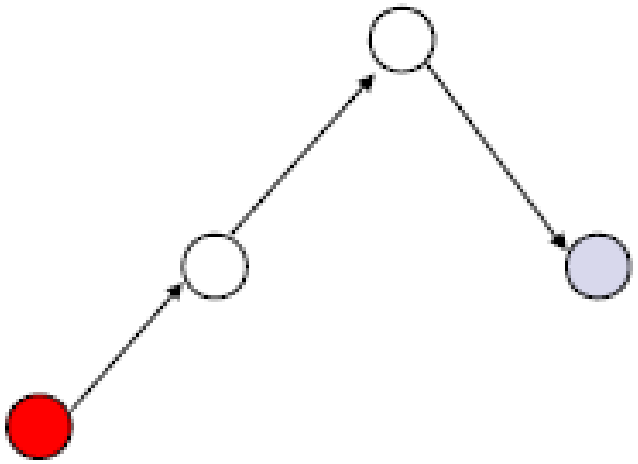
Compute Clear

Show version info

Created by Ted Pedersen and Jason Michelizzi
E-mail: tpederse (at) d (dot) umn (dot) edu

Đo quan hệ từ vựng

- Đếm số cạnh/đỉnh trên đồ thị:
  - khoảng cách giữa 2 từ tỉ lệ nghịch với quan hệ ngữ nghĩa giữa chúng
  - Nếu giữa 2 từ có nhiều đường đi, chọn đường ngắn nhất



số cạnh = 3

số nút = 4

Cặp từ nào gần nhau hơn?

- cá heo và cá?
- cá và cá hồi?

WordNet Similarity Metrics:

http://marimba.d.umn.edu/cgi-bin/similarity/similarity.cgi