Semantics

Semantics : Introduction

What is Semantics?

The study of meaning

- Relation between symbols and their denotata.

For eg.-

- John told Mary that the train moved out of the station at 3 o’clock.

Two Methodologies:

1. Lexical Semantics
2. Distributional Semantics

Lexical Semantics

Definition:

- **Lexical semantics is concerned with the systematic meaning related connections among lexical items**, and the internal meaning-related structure of individual lexical items.
- To identify the semantics of lexical items, we need to focus on the notion of **lexeme, an individual entry in the lexicon.**

What is a lexeme?

Lexeme should be thought of as a pairing of a particular orthographic and phonological form with some sort of symbolic meaning representation.

- Orthographic form, and phonological form refer to the appropriate form part of a lexeme
- Sense refers to a lexeme's meaning counterpart.

WordNet Dictionary: Example for word Verge

verge [1] | vərj |

noun
an edge or border: *they came down to the verge of the lake.*
• an extreme limit beyond which something specified will happen: *I was* ***on the verge of*** *tears.*
• Brit. a grass edging such as that by the side of a road or path.
• Architecture an edge of tiles projecting over a gable.

verb [no obj.] (**verge on**)
approach (something) closely; be close or similar to (something): *despair verging on the suicidal.*

ORIGIN late Middle English: via Old French from Latin ***virga 'rod.'*** The current verb sense dates from the late 18th cent.

verge [2] | vərj |

noun
a wand or rod carried before a bishop or dean as an emblem of office.

ORIGIN late Middle English: from Latin ***virga 'rod.'***

verge [3] | vərj |

verb [no obj.]
incline in a certain direction or toward a particular state: *his style verged into the art nouveau school.*

ORIGIN early 17th cent. (in the sense ***'descend (to the horizon)'***): from Latin ***vergere 'to bend, incline.'***

Relations between Word/ Lexeme meanings

1. Homonymy
2. Polysemy
3. Synonymy
4. Antonymy
5. Hypernymy
6. Hyponymy
7. Meronymy

Homonymy

Definition :

- **Homonymy** is defined as a relation that holds between words that have the same form with unrelated meanings.

Examples:

- Bat (wooden stick-like thing) vs Bat (flying mammal thing)
- Bank (financial institution) vs Bank (riverside)

Homophones and Homographs:

- **Homophones** are the words with the same pronunciation but different spellings.
  - write vs right
  - piece vs peace
- **Homographs** are the lexemes with the same orthographic form but different meaning.
  - Ex: bass (fish or Guitar)

    bank (financial institution or river bank)

Polysemy

- **Multiple related meanings within a single lexeme**.
  - The **bank** was constructed in 1875 out of local red brick.
  - I withdrew the money from the **bank**.
- Are **those** of the same sense?
  - **Sense 1:** "The building belonging to a financial institution"
  - **Sense 2:** "A financial institution"

Another example

  - Heavy snow caused the roof of **the school** to collapse.
  - **The school** hired more teachers this year than ever before.

Polysemy: building vs. organization

More examples:

- Tree (**Plums** have beautiful blossoms) ⇿ Fruit (I ate a preserved **plum** yesterday)
- Author (**J. K. Rowling** wrote Harry Potter series) ⇿ Works of Author (I really love **J. K. Rowling**)
- Animal (**The chicken** was domesticated in Asia) ⇿ Meat (**The chicken** was overcooked)

Synonymy

Words that have the same meaning in some or all contexts.

- filbert / hazelnut
- couch / sofa
- big / large
- automobile / car
- vomit / throw up
- water / H_2O

- Two lexemes are synonyms if they can be successfully substituted for each other in all situations.

Synonymy: A relation between senses

Consider the words **big** and **large**.

Are they synonyms?

- How **big** is that plane?
- Would I be flying on a **large** or small plane?

Synonymy: A relation between senses

Consider the words **big** and **large**.

Are they synonyms?

- How **big** is that plane?
- Would I be flying on a **large** or small plane?

How about here?

- Miss Nelson became a kind of **big sister** to Benjamin.
- Miss Nelson became a kind of **large sister** to Benjamin.

Why?

- **big** has a sense that means being older, or grown up
- **large** lacks this sense

Antonyms

- Senses that are opposites with respect to one feature of their meaning
- Otherwise, they are similar
  - dark / light
  - short / long
  - hot / cold
  - up / down
  - in / out

More formally: antonyms can

- define a binary opposition or at opposite ends of a scale (long/short, fast/slow)
- Be **reversive: rise/fall**

Hyponymy and Hypernymy

Hyponymy :

- One sense is a **hyponym** of another if the first sense is more specific, denoting a subclass of the other.
  - car is a hyponym of vehicle
  - dog is a hyponym of animal
  - mango is a hyponym of fruit

Hypernym/Superordinate(“hyper is super”)

Conversely,

- vehicle is a **hypernym/superordinate** of car
- animal is a **hypernym** of dog
- fruit is a **hypernym** of mango

Hyponymy

- **Extensional:**
  - The class denoted by the superordinate extensionally includes the class denoted by the hyponym
- **Entailment:**
  - A sense A is a hyponym of sense B if *being an A entails being a B*

  Eg. dog, animal
- **Hyponymy is usually transitive**
  - A hypo B and B hypo C entails A hypo C
- Another name: the **IS-A** hierarchy
  - A **IS-A** B **(or** A **ISA** B)
  - B subsumes A

WordNet Hierarchy :

Synonyms/ hypernyms (ordered by estimated frequency) of noun bass

8 senses of bass

Sense 1
bass -- (the lowest part of the musical range)
=> pitch -- (the property of sound that varies with variation in the frequency of vibration)
=> sound property -- (an attribute of sound)
=> property -- (a basic or essential attribute shared by all members of a class; "a study of the physical properties of atomic particles")
=> attribute -- (an abstraction belonging to or characteristic of an entity)
=> abstraction -- (a general concept formed by extracting common features from specific examples)
=> abstract entity -- (an entity that exists only abstractly)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Sense 2
bass, bass part -- (the lowest part in polyphonic music)
=> part, voice -- (the melody carried by a particular voice or instrument in polyphonic music; "he tried to sing the tenor part")
=> tune, melody, air, strain, melodic line, line, melodic phrase -- (a succession of notes forming a distinctive sequence; "she was humming an air from Beethoven")
=> music -- (an artistic form of auditory communication incorporating instrumental or vocal tones in a structured and continuous manner)
=> auditory communication -- (communication that relies on hearing)
=> communication -- (something that is communicated by or to or between people or groups)
=> abstraction -- (a general concept formed by extracting common features from specific examples)
=> abstract entity -- (an entity that exists only abstractly)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

WordNet Hierarchy:

Sense 3
bass, basso -- (an adult male singer with the lowest voice)
=> singer, vocalist, vocalizer, vocaliser -- (a person who sings)
=> musician, instrumentalist, player -- (someone who plays a musical instrument (as a profession))
=> performer, performing artist -- (an entertainer who performs a dramatic or musical work for an audience)
=> entertainer -- (a person who tries to please or amuse)
=> person, individual, someone, somebody, mortal, soul -- (a human being; "there was too much for one person to do")
=> organism, being -- (a living thing that has (or can develop) the ability to act or function independently)
=> living thing, animate thing -- (a living (or once living) entity)
=> object, physical object -- (a tangible and visible entity; an entity that can cast a shadow; "it was full of rackets, balls and other objects")
=> physical entity -- (an entity that has physical existence)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
=> causal agent, cause, causal agency -- (any entity that produces an effect or is responsible for events or results)
=> physical entity -- (an entity that has physical existence)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

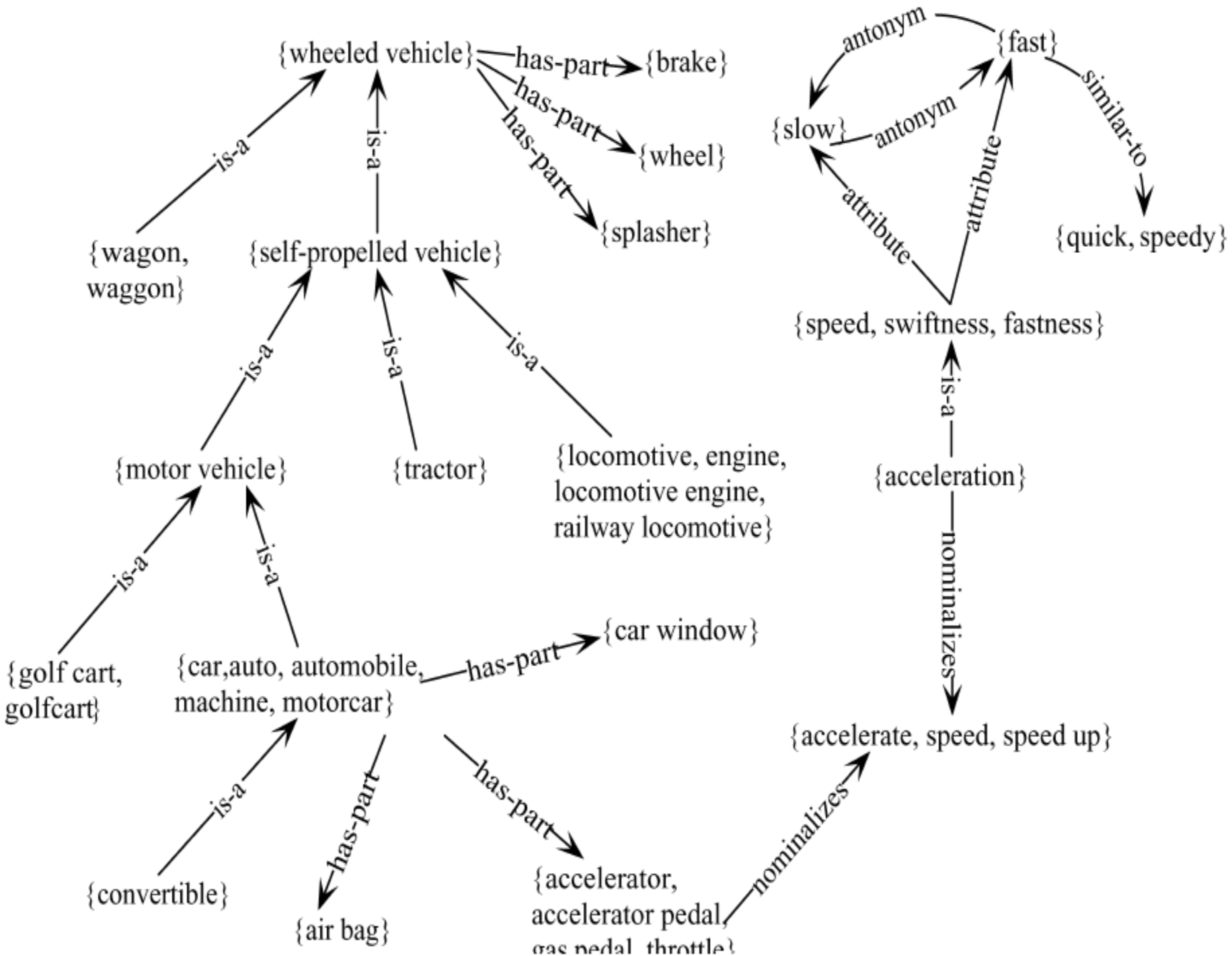
Sense 4
sea bass, bass -- (the lean flesh of a saltwater fish of the family Serranidae)
=> saltwater fish -- (flesh of fish from the sea used as food)
=> seafood -- (edible fish (broadly including freshwater fish) or shellfish or roe etc)
=> food, solid food -- (any solid substance (as opposed to liquid) that is used as a source of nourishment; "food and drink")
=> solid -- (a substance that is solid at room temperature and pressure)
=> substance, matter -- (that which has mass and occupies space; "an atom is the smallest indivisible unit of matter")
=> physical entity -- (an entity that has physical existence)
=> entity -- (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Meronyms and Holonyms:

Meronymy: an asymmetric, transitive relation between senses.

- X is a **meronym** of Y if X denotes **a part** of Y.
- The inverse relation is **holonymy.**
- The part-whole relation
  - a *leg is **part of** a chair*
  - *a wheel is **part o**f a car*
- *Wheel is a **meronym** of car,*
- *Car is a **holonym** of wheel.*

meronym	**holonym**
porch	House
nose	face
wheel	Car
Leg	chair


{wheeled vehicle}
has-part
{brake}
has-part
{wheel}
has-part
{splasher}
is-a
is-a
{wagon, waggon}
{self-propelled vehicle}
is-a
is-a
is-a
{motor vehicle}
{tractor}
{locomotive, engine, locomotive engine, railway locomotive}
is-a
is-a
{golf cart, golfcart}
{car,auto, automobile, machine, motorcar}
has-part
{car window}
is-a
has-part
has-part
{convertible}
{air bag}
{accelerator, accelerator pedal, gas pedal, throttle}
nominalizes
{accelerate, speed, speed up}
nominalizes
{acceleration}
is-a
{speed, swiftness, fastness}
attribute
attribute
{slow}
antonym
antonym
{fast}
similar-to
{quick, speedy}

WordNet: https://wordnet.princeton.edu/wordnet/

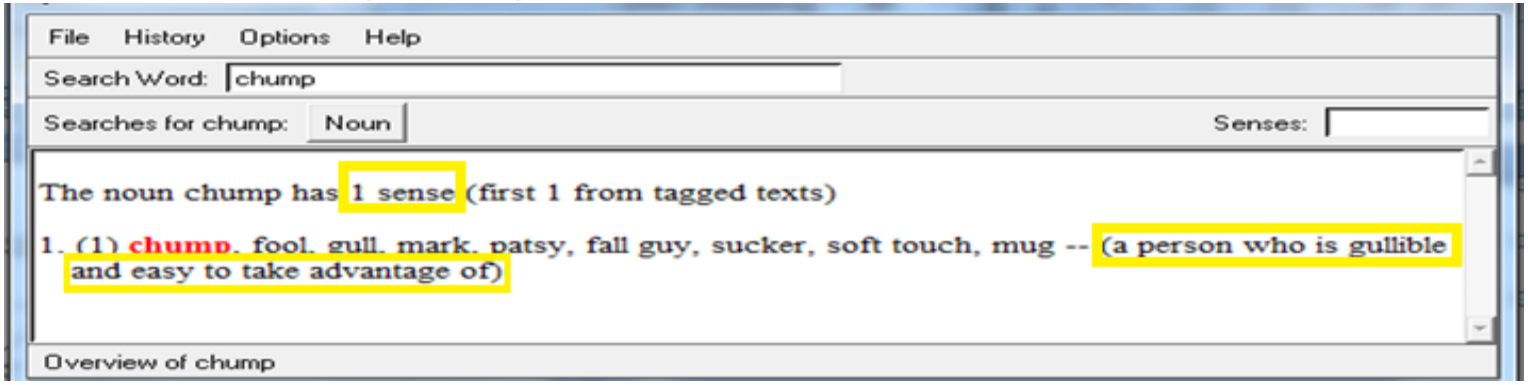
- In 1993, WordNet was introduced.
- It is a hierarchically organized lexical database, organized as a semantic network.
- The development began in 1985 at Princeton University by a group of psychologists and linguists,
- A machine-readable thesaurus
- WordNet consists of four separate databases, for nouns, verbs, adjectives and adverbs.
- The WordNet 3.0 released in December 2006.
- It contains:
  - 117,7987 nouns,
  - 22,479 adjectives,
  - 11,529 verbs ,and
  - 4,481 adverbs.

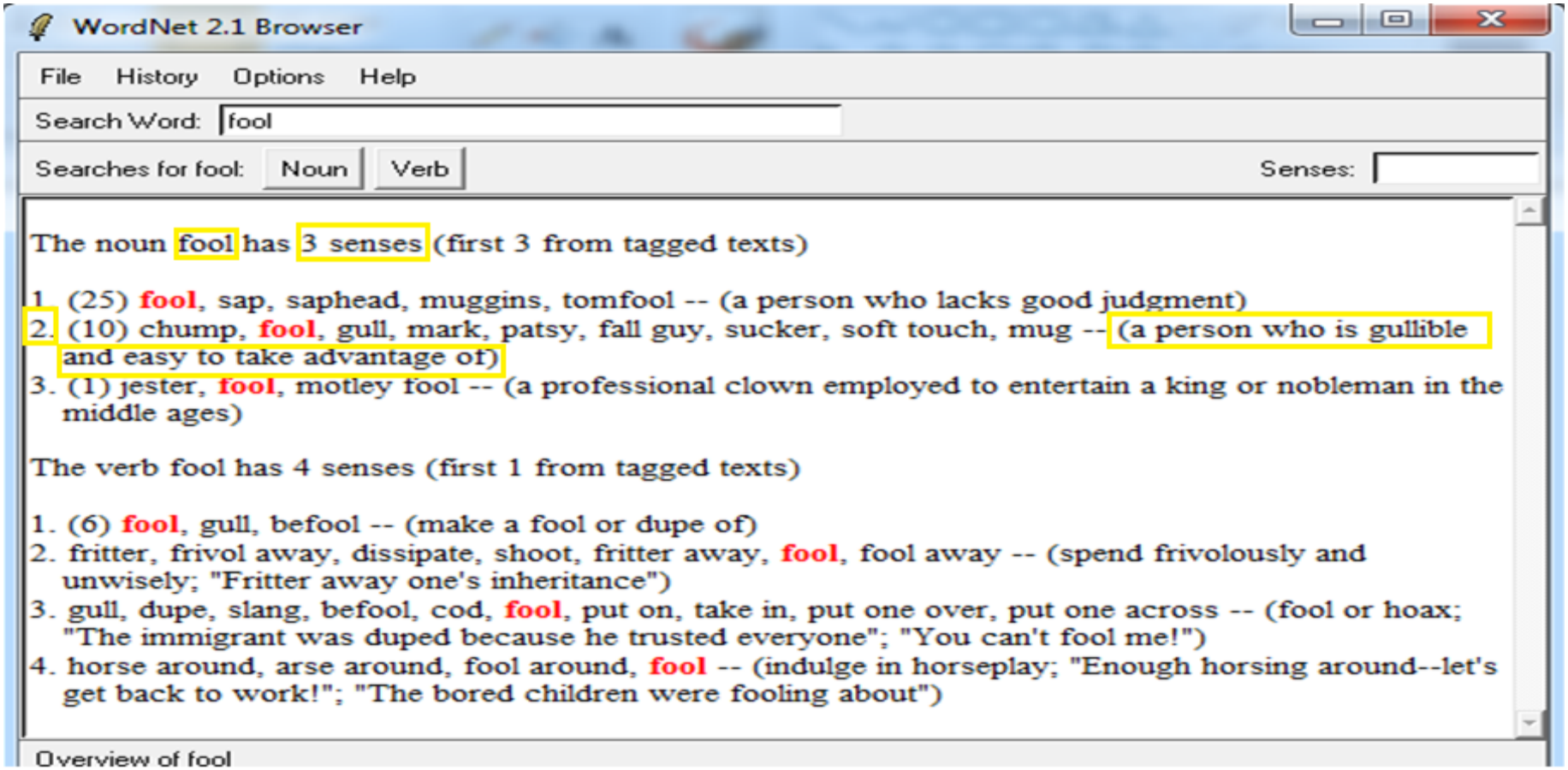
Synsets in WordNet

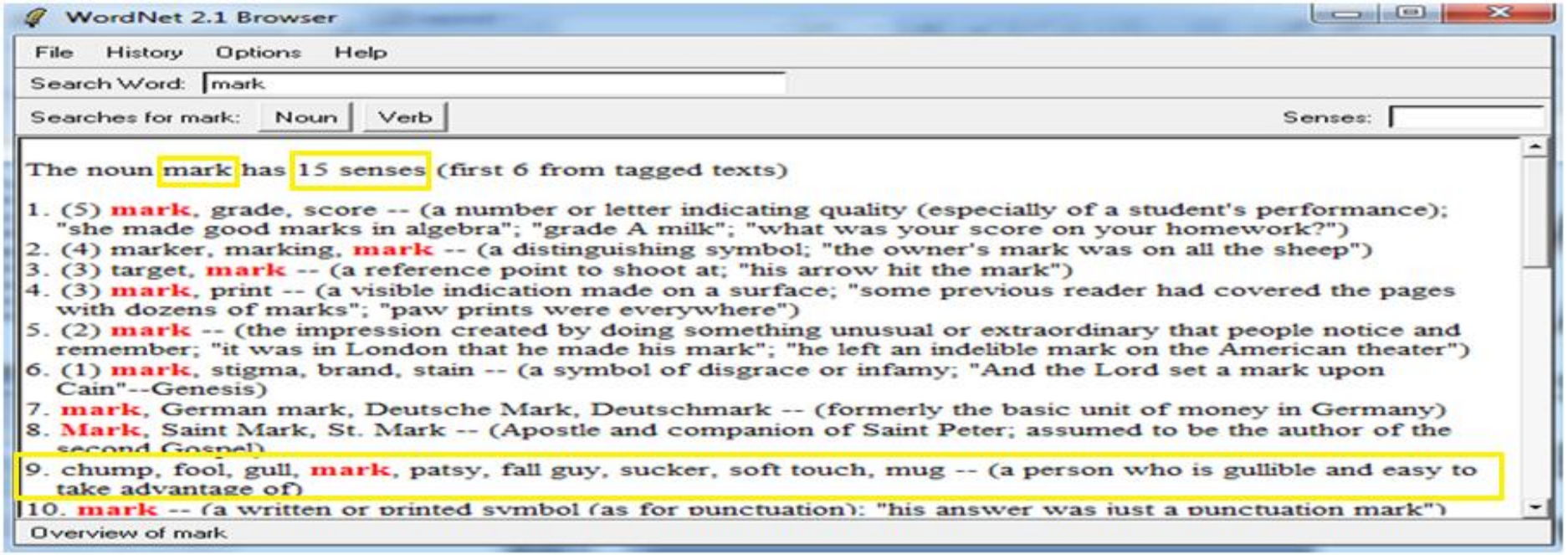
- A **synset** is a set of synonyms representing a sense.
- Example: **chump** as a noun to mean '**a person who is gullible and easy to take advantage of**'

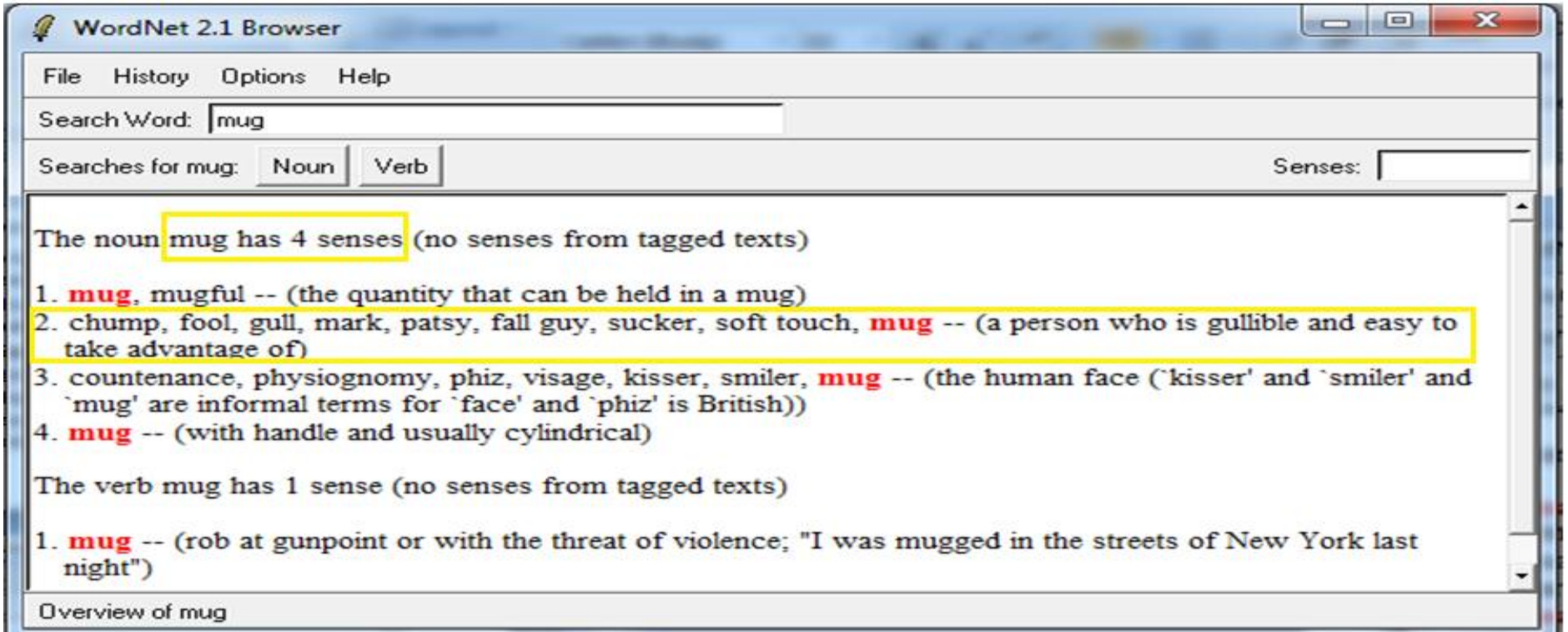
  Synset: {chump^{1}, fool^{2}, gull^{1}, mark^{9}, patsy^{1}, fall guy^{1}, sucker^{1}, soft touch^{1}, mug^{2}}
- Each of these senses , like sense1 of chump, sense2 of fool, …. , share the same gloss definition.

Chump Synsets

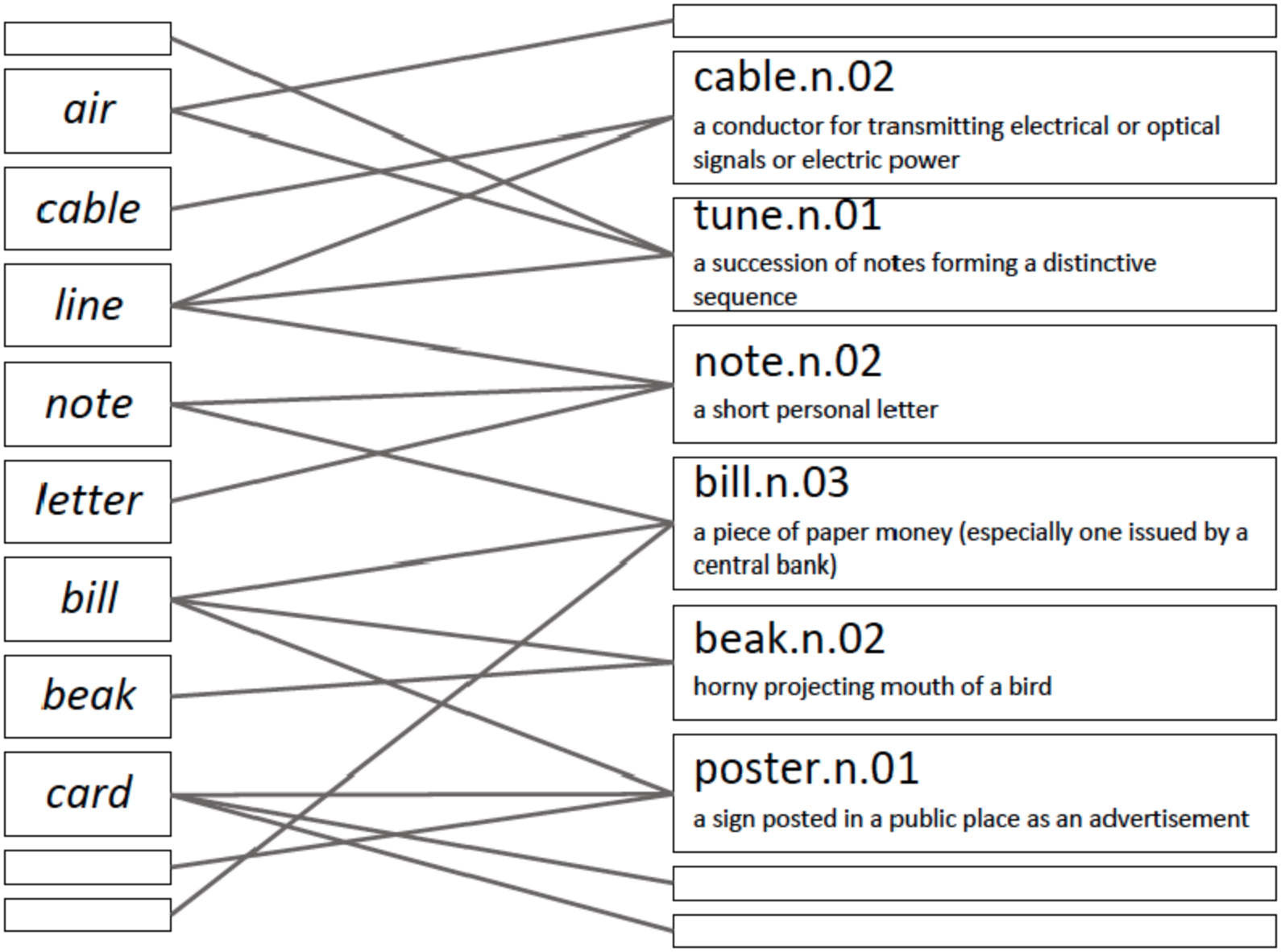





WordNet 2.1 Browser
File History Options Help
Search Word: mark
Searches for mark: Noun Verb
Senses:
The noun mark has 15 senses (first 6 from tagged texts)
1. (5) mark, grade, score -- (a number or letter indicating quality (especially of a student's performance); "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?")
2. (4) marker, marking, mark -- (a distinguishing symbol; "the owner's mark was on all the sheep")
3. (3) target, mark -- (a reference point to shoot at; "his arrow hit the mark")
4. (3) mark, print -- (a visible indication made on a surface; "some previous reader had covered the pages with dozens of marks"; "paw prints were everywhere")
5. (2) mark -- (the impression created by doing something unusual or extraordinary that people notice and remember; "it was in London that he made his mark"; "he left an indelible mark on the American theater")
6. (1) mark, stigma, brand, stain -- (a symbol of disgrace or infamy; "And the Lord set a mark upon Cain"--Genesis)
7. mark, German mark, Deutsche Mark, Deutschmark -- (formerly the basic unit of money in Germany)
8. Mark, Saint Mark, St. Mark -- (Apostle and companion of Saint Peter; assumed to be the author of the second Gospel)
9. chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug -- (a person who is gullible and easy to take advantage of)
10. mark -- (a written or printed symbol (as for punctuation); "his answer was just a punctuation mark")
Overview of mark


WordNet 2.1 Browser
File History Options Help
Search Word: mug
Searches for mug: Noun Verb
Senses:
The noun mug has 4 senses (no senses from tagged texts)
1. mug, mugful -- (the quantity that can be held in a mug)
2. chump, fool, gull, mark, patsy, fall guy, sucker, soft touch, mug -- (a person who is gullible and easy to take advantage of)
3. countenance, physiognomy, phiz, visage, kisser, smiler, mug -- (the human face (`kisser' and `smiler' and `mug' are informal terms for `face' and `phiz' is British))
4. mug -- (with handle and usually cylindrical)
The verb mug has 1 sense (no senses from tagged texts)
1. mug -- (rob at gunpoint or with the threat of violence; "I was mugged in the streets of New York last night")
Overview of mug

lemma vs. Synsets



WordNet Hierarchies:

```
Synonyms/Hypernyms (Ordered by Estimated Frequency) of noun mouse

4 senses of mouse

Sense 1
mouse
        => rodent, gnawer
            => placental, placental mammal, eutherian, eutherian mammal
                => mammal, mammalian
                    => vertebrate, craniate
                        => chordate
                            => animal, animate being, beast, brute, creature, fauna
                                => organism, being
                                    => living thing, animate thing
                                        => whole, unit
                                            => object, physical object
                                                => physical entity
                                                    => entity


Sense 4
mouse, computer mouse
        => electronic device
            => device
                => instrumentality, instrumentation
                    => artifact, artefact
                        => whole, unit
                            => object, physical object
                                => physical entity
                                    => entity
```

All relations in WordNet

```
searchtype is at least one of the following:
        -ants{n|v|a|r}          Antonyms
        -hype{n|v}              Hypernyms
        -hypo{n|v}, -tree{n|v}  Hyponyms & Hyponym Tree
        -entav                  Verb Entailment
        -syns{n|v|a|r}          Synonyms (ordered by estimated frequency)
        -smemn                  Member of Holonyms
        -ssubn                  Substance of Holonyms
        -sprtn                  Part of Holonyms
        -membn                  Has Member Meronyms
        -subsn                  Has Substance Meronyms
        -partn                  Has Part Meronyms
        -meron                  All Meronyms
        -holon                  All Holonyms
        -causv                  Cause to
        -pert{a|r}              Pertainyms
        -attr{n|a}              Attributes
        -deri{n|v}              Derived Forms
        -domn{n|v|a|r}          Domain
        -domt{n|v|a|r}          Domain Terms
        -faml{n|v|a|r}          Familiarity & Polysemy Count
        -framv                  Verb Frames
        -coor{n|v}              Coordinate Terms (sisters)
        -simsv                  Synonyms (grouped by similarity of meaning)
        -hmern                  Hierarchical Meronyms
        -hholn                  Hierarchical Holonyms
        -grep{n|v|a|r}          List of Compound Words
        -over                   Overview of Senses
```

Word Similarity

- Synonymy is a binary relation
  - Two words are either synonymous or not
- Two words are more similar
  - If they share more features of meaning
- Actually these are relations between **senses:**
- Instead of saying **"bank is like fund"**
- We say:
  - **Bank^{8}** is similar to **fund^{1}**
  - **Bank^{2}** is similar to **slope^{1}**
- We will compute similarity over both: words and senses

Bank and Fund WordNet Definitions:

Wordnet Dictionary | Home | About | Contact | Test English Info | Google Dictionary

Wordnet Dictionary

Search

bank

Noun bank has 10 senses

1. depository financial institution, bank, banking concern, banking company - a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"
   - --$_1$ is a kind of financial institution, financial organization, financial organisation
   - --$_1$ is a member of banking industry, banking system
   - --$_1$ has particulars:
     credit union; Federal Reserve Bank, reserve bank; agent bank; commercial bank, full service bank; state bank; lead bank, agent bank; member bank; merchant bank, acquirer; acquirer; thrift institution; Home Loan Bank
   - Derived forms: verb bank$_3$, verb bank$_5$, verb bank$_6$
2. bank - sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"
   - --$_2$ is a kind of slope, incline, side
   - --$_2$ has particulars: riverbank, riverside; waterside
   - Derived form: verb bank$_2$
3. bank - a supply or stock held in reserve for future use (especially in emergencies)
   - --$_3$ is a kind of reserve, backlog, stockpile
   - --$_3$ has particulars: blood bank; eye bank; food bank; soil bank
4. bank, bank building - a building in which commercial banking is transacted; "the bank is on the corner of Nassau and Witherspoon"
   - --$_4$ is a kind of depository, deposit, repository
   - --$_4$ has parts: vault, bank vault
   - Derived forms: verb bank$_3$, verb bank$_5$, verb bank$_6$
5. bank - an arrangement of similar objects in a row or in tiers; "he operated a bank of switches"
   - --$_5$ is a kind of array
6. savings bank, coin bank, money box, bank - a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty"
   - --$_6$ is a kind of container
   - --$_6$ has particulars: piggy bank, penny bank
7. bank - a long ridge or pile; "a huge bank of earth"
   - --$_7$ is a kind of ridge
   - --$_7$ has particulars: bluff; sandbank
8. bank - the funds held by a gambling house or the dealer in some gambling games; "he tried to break the bank at Monte Carlo"
   - --$_8$ is a kind of funds, finances, monetary resource, cash in hand, pecuniary resource
9. bank, cant, camber - a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force
   - --$_9$ is a kind of slope, incline, side
10. bank - a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank"
    - --$_{10}$ is a kind of flight maneuver, airplane maneuver
    - --$_{10}$ has particulars: vertical bank
    - Derived form: verb bank$_1$

Verb bank has 8 senses

1. bank - tip laterally; "the pilot had to bank the aircraft"
   - --$_1$ is one way to tip

Wordnet Dictionary | Home | About | Contact | Test English Info | Google Dictionary

Wordnet Dictionary

Search

fund

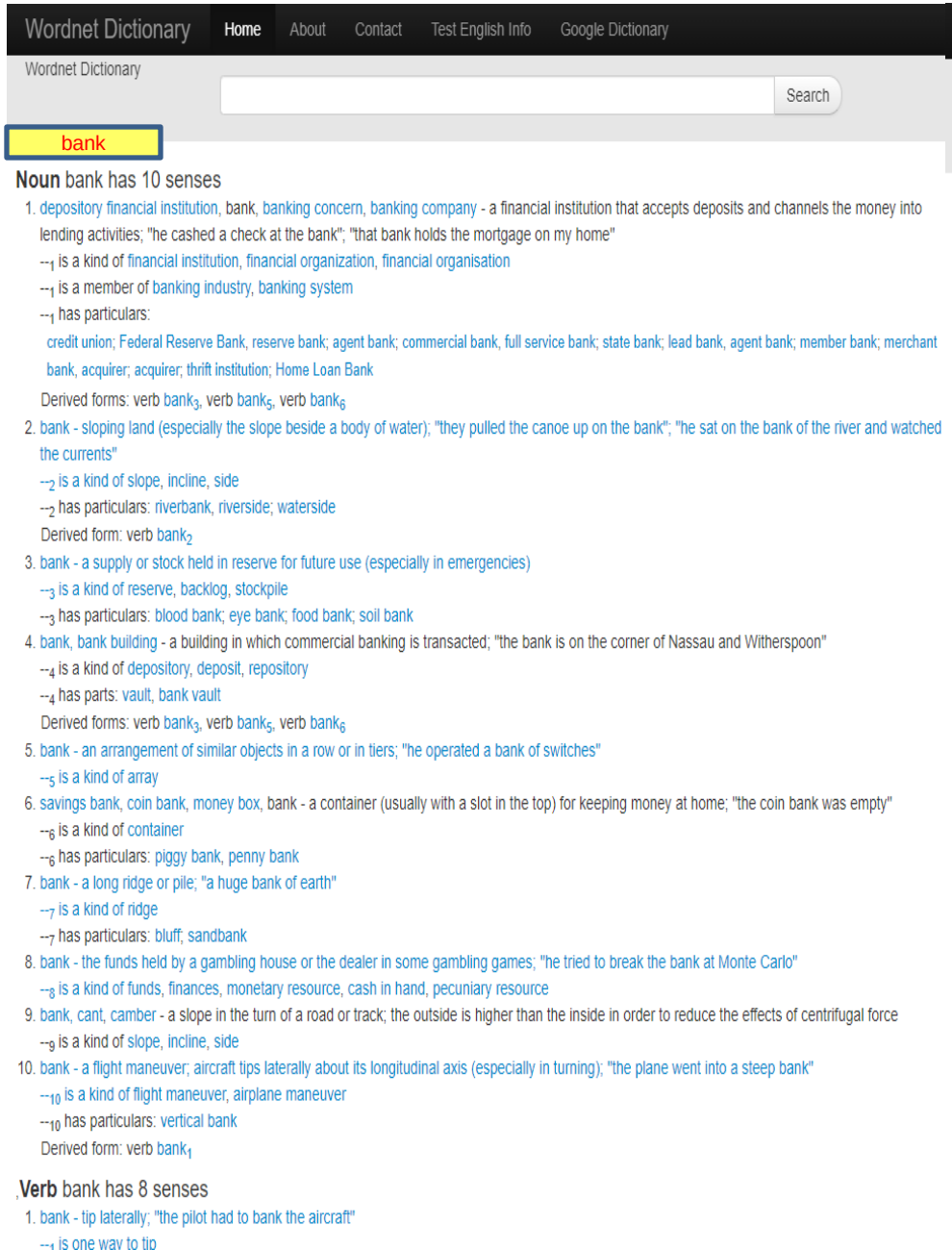
Noun fund has 3 senses

1. fund, monetary fund - a reserve of money set aside for some purpose
   - --$_1$ is a kind of money
   - --$_1$ has particulars:
     revolving fund; sinking fund; savings, nest egg; pension fund, superannuation fund; war chest; slush fund; trust fund; deposit, bank deposit; budget; petty cash
   - Derived forms: verb fund$_6$, verb fund$_2$
2. store, stock, fund - a supply of something available for future use; "he brought back a large store of Cuban cigars"
   - --$_2$ is a kind of accumulation
   - --$_2$ has particulars:
     infrastructure, base; hoard, cache, stash; provision; issue, military issue, government issue; seed stock
3. investment company, investment trust, investment firm, fund - a financial institution that sells shares to individuals and invests in securities issued by other companies
   - --$_3$ is a kind of nondepository financial institution
   - --$_3$ has particulars:
     hedge fund, hedgefund; mutual fund, open-end fund, open-end investment company; closed-end fund, closed-end investment company; face-amount certificate company; Real Estate Investment Trust, REIT; unit investment trust, unit trust; nondiscretionary trust, fixed investment trust
   - Derived forms: verb fund$_4$, verb fund$_3$

Verb fund has 6 senses

1. fund - convert (short-term floating debt) into long-term debt that bears fixed interest and is represented by bonds
   - --$_1$ is one way to finance
   - Sample sentence:
   - Somebody ----s something
2. fund - place or store up in a fund for accumulation
   - --$_2$ is one way to roll up, collect, accumulate, pile up, amass, compile, hoard
   - Derived form: noun fund$_1$
   - Sample sentence:
   - Somebody ----s something
3. fund - provide a fund for the redemption of principal or payment of interest
   - --$_3$ is one way to supply, provide, render, furnish
   - Derived form: noun fund$_3$
   - Sample sentence:
   - Somebody ----s something
4. fund - invest money in government securities
   - --$_4$ is one way to invest, put, commit, place
   - Derived form: noun fund$_3$
   - Sample sentence:
   - Somebody ----s something
5. fund - accumulate a fund for the discharge of a recurrent liability; "fund a medical care plan"
   - --$_5$ is one way to roll up, collect, accumulate, pile up, amass, compile, hoard

Bank and Slope WordNet Definitions:

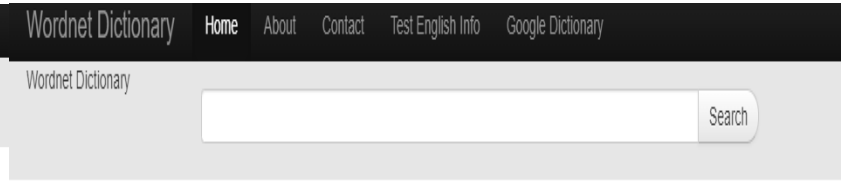


Noun bank has 10 senses

1. depository financial institution, bank, banking concern, banking company - a financial institution that accepts deposits and channels the money into lending activities; "he cashed a check at the bank"; "that bank holds the mortgage on my home"
   - --$_1$ is a kind of financial institution, financial organization, financial organisation
   - --$_1$ is a member of banking industry, banking system
   - --$_1$ has particulars:
     credit union; Federal Reserve Bank, reserve bank; agent bank; commercial bank, full service bank; state bank; lead bank, agent bank; member bank; merchant bank, acquirer; acquirer; thrift institution; Home Loan Bank
   - Derived forms: verb bank$_3$, verb bank$_5$, verb bank$_6$
2. bank - sloping land (especially the slope beside a body of water); "they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"
   - --$_2$ is a kind of slope, incline, side
   - --$_2$ has particulars: riverbank, riverside; waterside
   - Derived form: verb bank$_2$
3. bank - a supply or stock held in reserve for future use (especially in emergencies)
   - --$_3$ is a kind of reserve, backlog, stockpile
   - --$_3$ has particulars: blood bank; eye bank; food bank; soil bank
4. bank, bank building - a building in which commercial banking is transacted; "the bank is on the corner of Nassau and Witherspoon"
   - --$_4$ is a kind of depository, deposit, repository
   - --$_4$ has parts: vault, bank vault
   - Derived forms: verb bank$_3$, verb bank$_5$, verb bank$_6$
5. bank - an arrangement of similar objects in a row or in tiers; "he operated a bank of switches"
   - --$_5$ is a kind of array
6. savings bank, coin bank, money box, bank - a container (usually with a slot in the top) for keeping money at home; "the coin bank was empty"
   - --$_6$ is a kind of container
   - --$_6$ has particulars: piggy bank, penny bank
7. bank - a long ridge or pile; "a huge bank of earth"
   - --$_7$ is a kind of ridge
   - --$_7$ has particulars: bluff; sandbank
8. bank - the funds held by a gambling house or the dealer in some gambling games; "he tried to break the bank at Monte Carlo"
   - --$_8$ is a kind of funds, finances, monetary resource, cash in hand, pecuniary resource
9. bank, cant, camber - a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force
   - --$_9$ is a kind of slope, incline, side
10. bank - a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning); "the plane went into a steep bank"
    - --$_{10}$ is a kind of flight maneuver, airplane maneuver
    - --$_{10}$ has particulars: vertical bank
    - Derived form: verb bank$_1$

Verb bank has 8 senses

1. bank - tip laterally; "the pilot had to bank the aircraft"
   - --$_1$ is one way to tip



Slope

Noun slope has 2 senses

1. slope, incline, side - an elevated geological formation; "he climbed the steep slope"; "the house was built on the side of the mountain"
   - --$_1$ is a kind of geological formation, formation
   - --$_1$ is a part of natural elevation, elevation
   - --$_1$ has particulars:
     ascent, acclivity, rise, raise, climb, upgrade; bank; bank, cant, camber; canyonside; coast; descent, declivity, fall, decline, declination, declension, downslope; escarpment, scarp; hillside; mountainside, versant; ski slope
   - Derived form: verb slope$_1$
2. gradient, slope - the property possessed by a line or surface that departs from the horizontal; "a five-degree gradient"
   - --$_2$ is a kind of position, spatial relation
   - --$_2$ has particulars:
     grade; pitch, rake, slant; abruptness, precipitousness, steepness; gradualness, gentleness
   - Derived form: verb slope$_1$

Verb slope has 1 sense

1. slope, incline, pitch - be at an angle; "The terrain sloped down"
   - --$_1$ is one way to lean, tilt, tip, slant, angle
   - Derived forms: noun slope$_2$, noun slope$_1$
   - Sample sentences:
   - Something ----s
   - Something is ----ing PP

Two classes of algorithms

1. **Distributional algorithms**
   - By comparing words based on their distributional context in the corpora
2. **Thesaurus-based algorithms**
   - Based on whether words are “nearby” in WordNet

Thesaurus-based Approaches

- **Thesaurus-based algorithms**
  - If two words are near in the hierarchical organization of a thesaurus, we say they are similar; if they are very far apart in the hierarchy, they might be different.
  - Using the WordNet resource, we will use this idea to establish similarity between two words.
  - We could use anything(relation) in the thesaurus:
    - Hypernymy, hyponymy, holonymy,…
    - Glosses and example sentences
  - In practice, “thesaurus based” measure usually use:
    - the **is-a**/hypernymy hierarchy
    - and sometimes the glosses too

Semantic similarity between words and word relatedness

- Semantic similarity or relatedness is the degree to which two concepts are related.
- Relatedness is more general than similarity, in that two concepts can be related although they are not similar.

For eg, rich and poor are related with antonymy relation in WordNet but they are not similar.

- Similarity measures are limited to the **is-a** hierarchies in WordNet whereas relatedness measures can be applied to all kinds of relations.
- Related words:
  - Car, gasoline : related, but not similar
  - Car, tire: related by automotive senses

Methodologies for Word Similarity

1. Path-based Similarity measure
2. Leacock-Chodorow (L-C) Similarity
3. Resnik Similarity
4. Lin Similarity
5. Jiang-Conrath Similarity

Path-based similarity

Basic Idea:

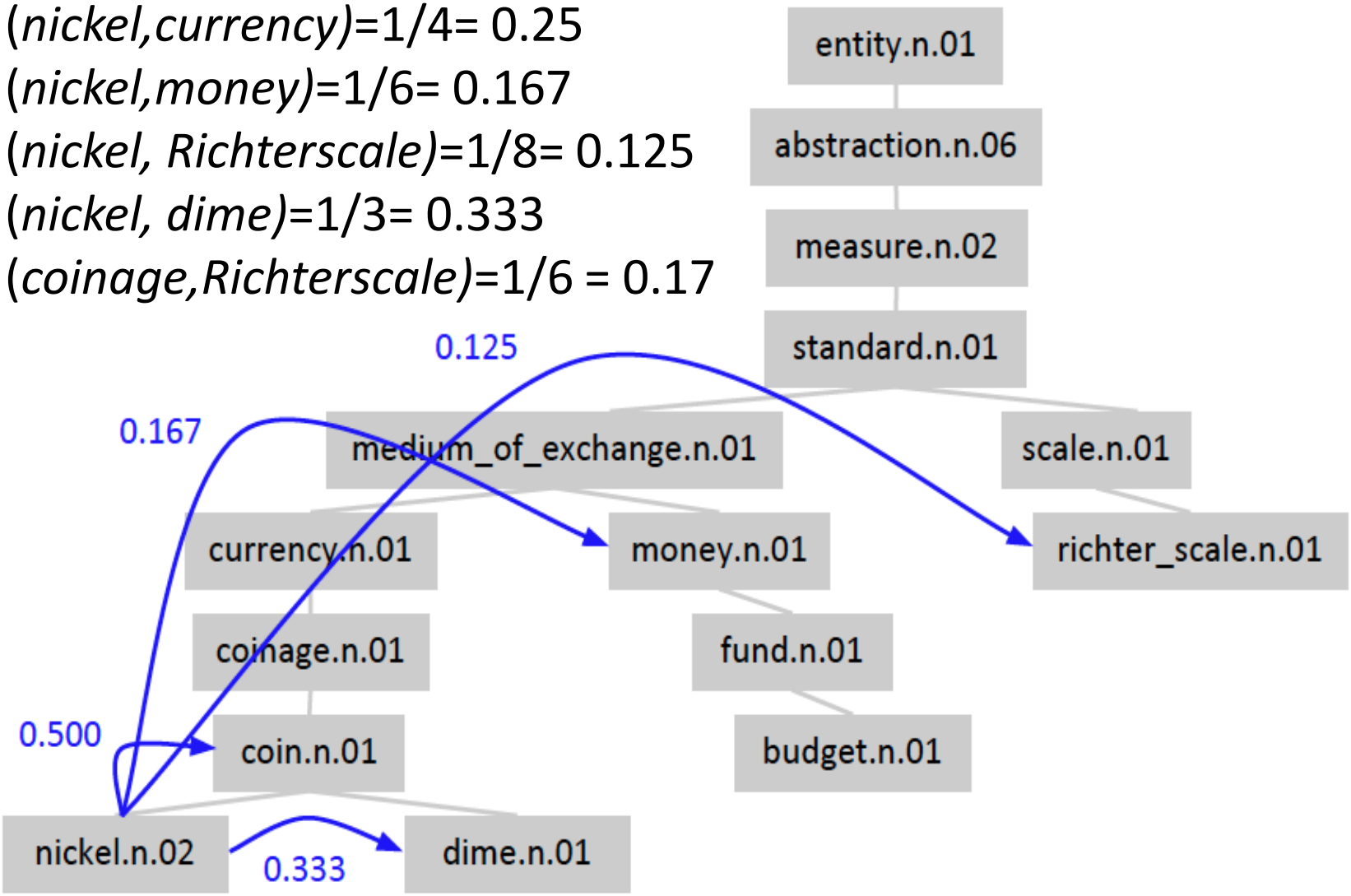
- Two words are similar if they are nearby in the hypernym graph
- pathlen(c_1, c_2) = number of edges in shortest path (in hypernym graph) between senses c_1 and c_2

$$sim_{path}(c_1, c_2) = \frac{1}{1 + pathlen(c_1, c_2)}$$

$$sim(w_1, w_2) = max_{c_1 \in senses(w_1), c_2 \in senses(w_2)} sim(c_1, c_2)$$

Shortest path in the hierarchy

sim_{path} (*nickel, coin)*=1/2 =0.5
sim_{path} (*fund,budget)*=1/2 =0.5
sim_{path} (*nickel,currency)*=1/4= 0.25
sim_{path} (*nickel,money)*=1/6= 0.167
sim_{path} (*nickel, Richterscale)*=1/8= 0.125
sim_{path} (*nickel, dime)*=1/3= 0.333
sim_{path} (*coinage,Richterscale)*=1/6 = 0.17



Leacock-Chodorow (L-C) Similarity

L-C similarity : The similarity between two concepts is measured using the path length between the concepts and then scaling it by the depth d of the taxonomy

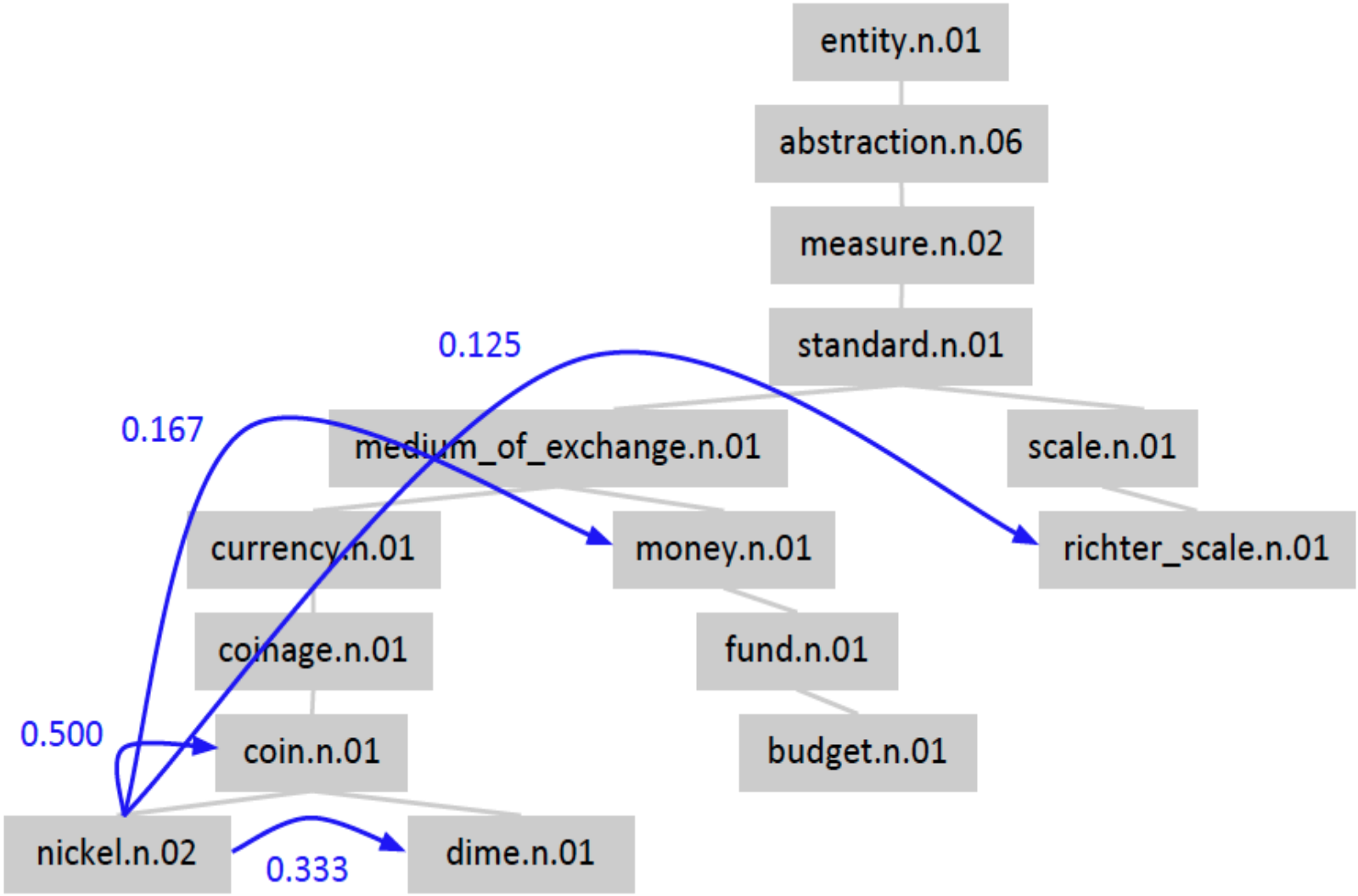
$$\mathbf{sim_{LC}(c_1, c_2) = -log(pathlen(c_1, c_2)/2d)}$$

where,

d: maximum depth of the hierarchy defined as longest path between leaf node and root of the taxonomy.

L-C similarity

$sim_{L\text{-}C}$ (*nickel, coin)*= -log(0.5/2*10) =-0.3979
$sim_{L\text{-}C}$ (*entity,abstraction)*=-log(0.5/2*10) =-0.3979



Problems with L-C similarity

- For any two pairs of concepts, if the path length is same, the similarity will be the same irrespective of wherever they occur in the tree.
- Eg:
  - similarity between coin and nickel = 0.5
  - similarity between entity and abstraction =0.5
- We are going down in the hierarchy, we are moving to very, very specific concepts.
- While we are moving to specific concepts, the same path length should amount to a higher similarity value that it was doing earlier.

Problems with L-C similarity

- We want a metric which lets us assign different “lengths” to different “edges”
- We want a metric that
  - represents the cost of each edge independently
  - words connected only through abstract nodes

Concept Probability models

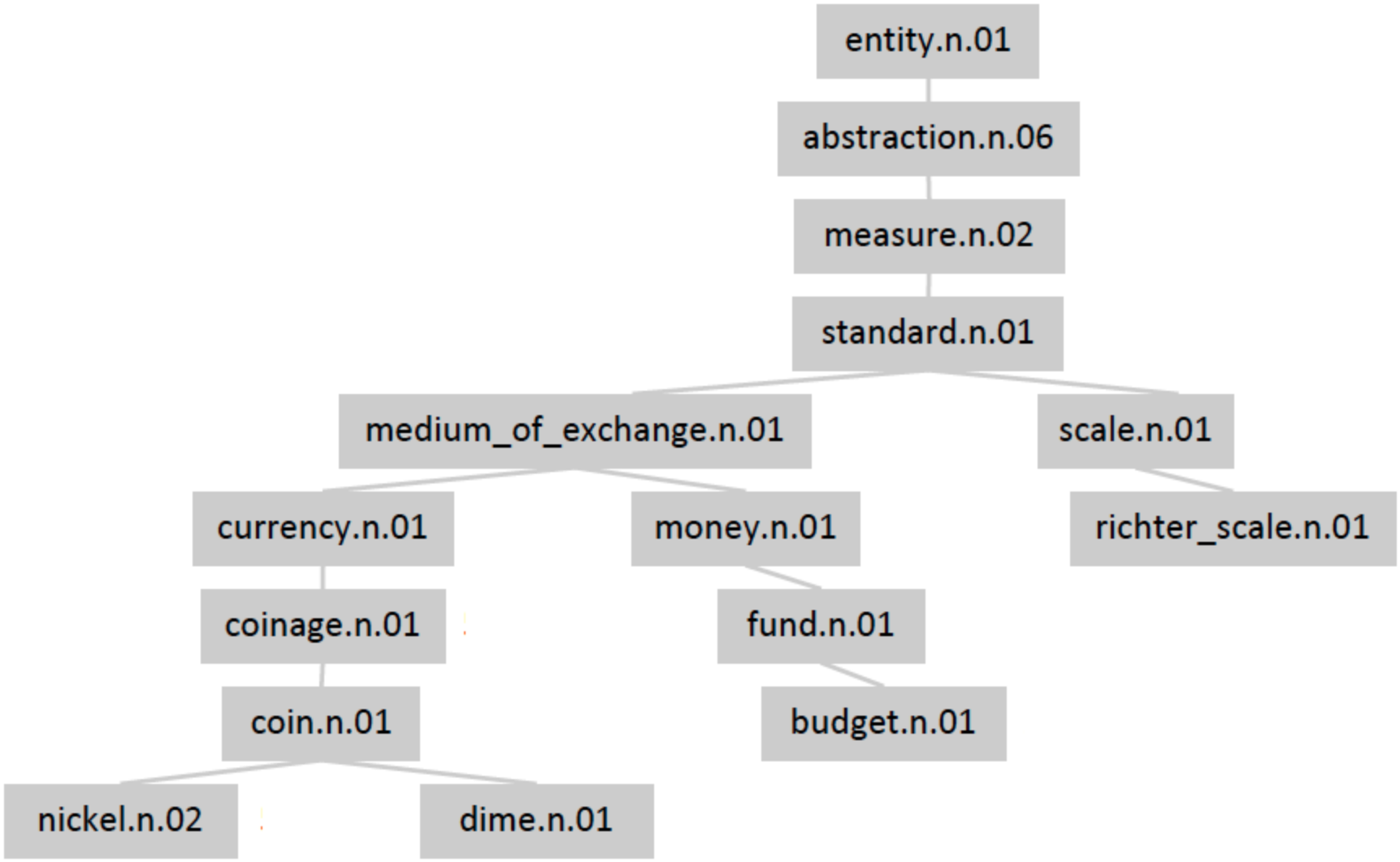
Concept probabilities:

- For each concept (synset) c, let P(c) be the probability of concept c in the hierarchy.
- **Idea would be**, whatever we are seeing in tree is an **entity**, because it is part of the tree where entity is the root. So, whatever word is in the tree is an **entity**.
- So, whenever a word is encounter, find out what are all the concepts to which it contributes, and **add a count of 1** to all these concepts.
- Finally, convert them to probability values.

Concept Probability models

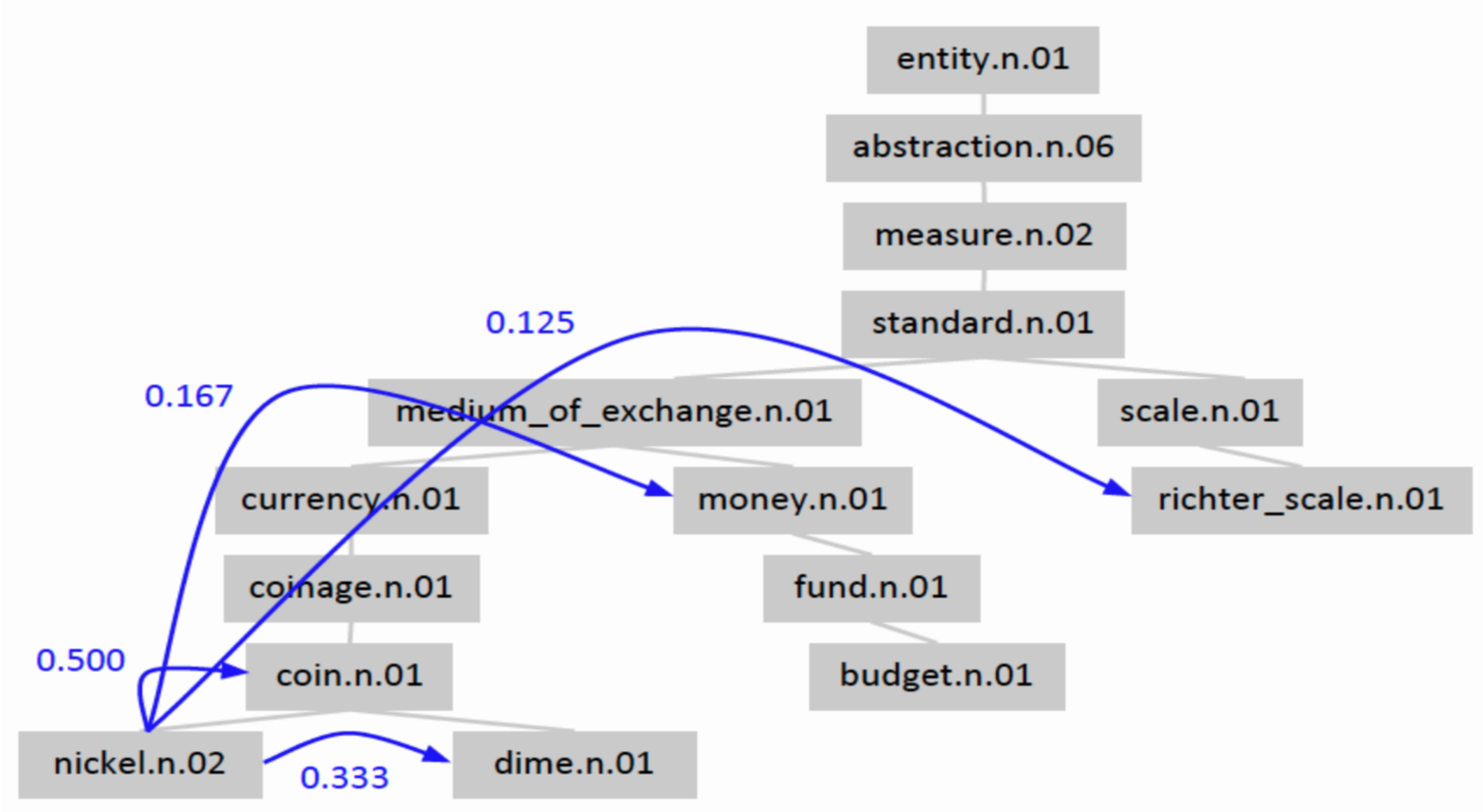
So, what would happen ?

- The root node will get a probability of 1, because everything we see is an entity, but as we go down these values will keep on decreasing.
  - P(Root)=1
  - The lower the node in hierarchy, lower its probability
- Use this idea to convert them into log values and then taking the difference between the two values as the path length which can be converted to finding the similarity between two synsets.

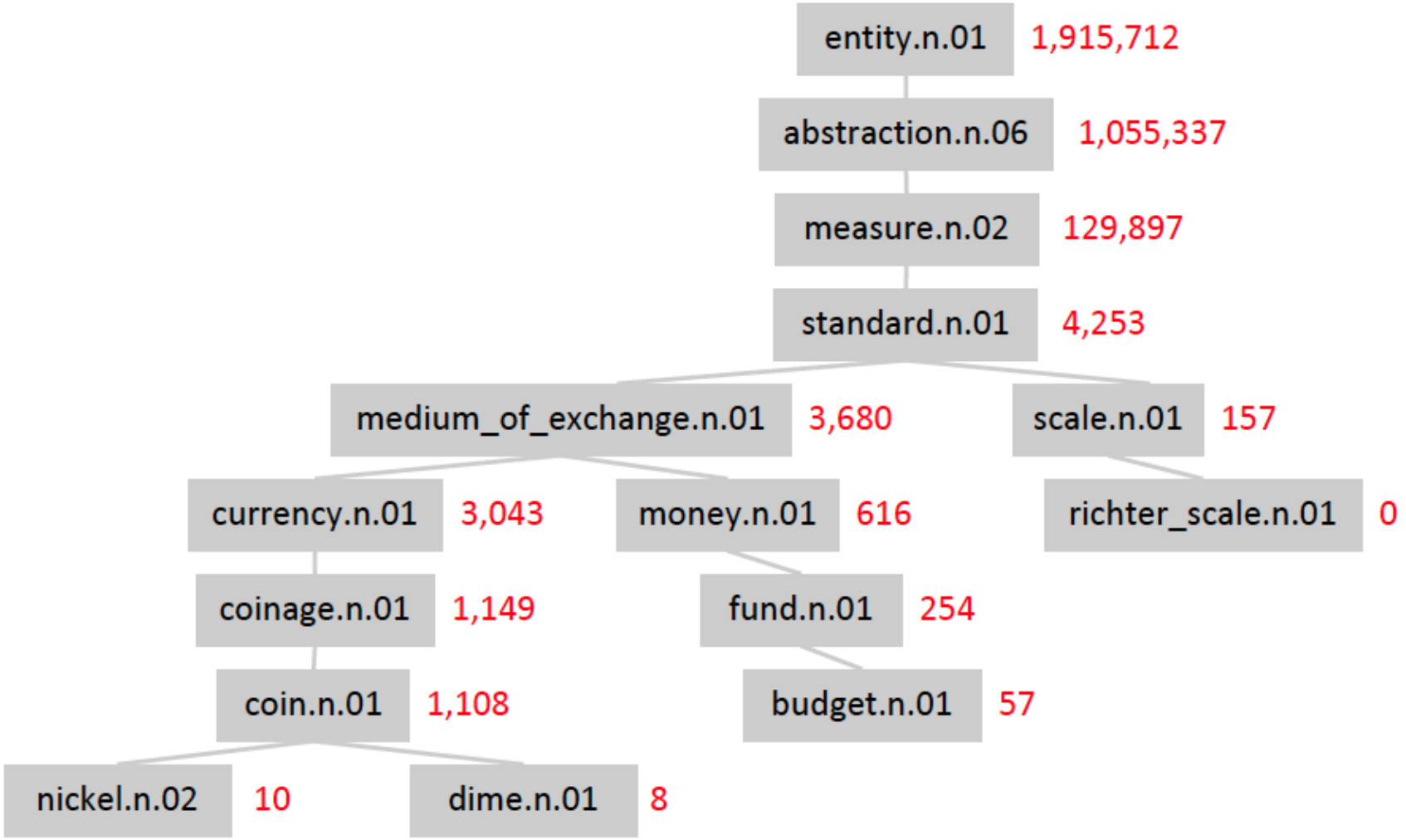

entity.n.01
abstraction.n.06
measure.n.02
standard.n.01
medium_of_exchange.n.01
scale.n.01
currency.n.01
money.n.01
richter_scale.n.01
coinage.n.01
fund.n.01
coin.n.01
budget.n.01
nickel.n.02
dime.n.01

Estimating Concept Probabilities

- Train by counting “concept activations” in a corpus
- Each occurrence of dime also increments counts for coin, coinage, currency, medium of exchange, standard, etc.



Example : Concept count



Each occurrence of dime also increments counts for coin, coinage, currency, medium_of_exchange, standard, etc

Example : Concept Probabilities

P(entity)=entity_count/N = **1**

entity_count= 1915712, N= 1915712

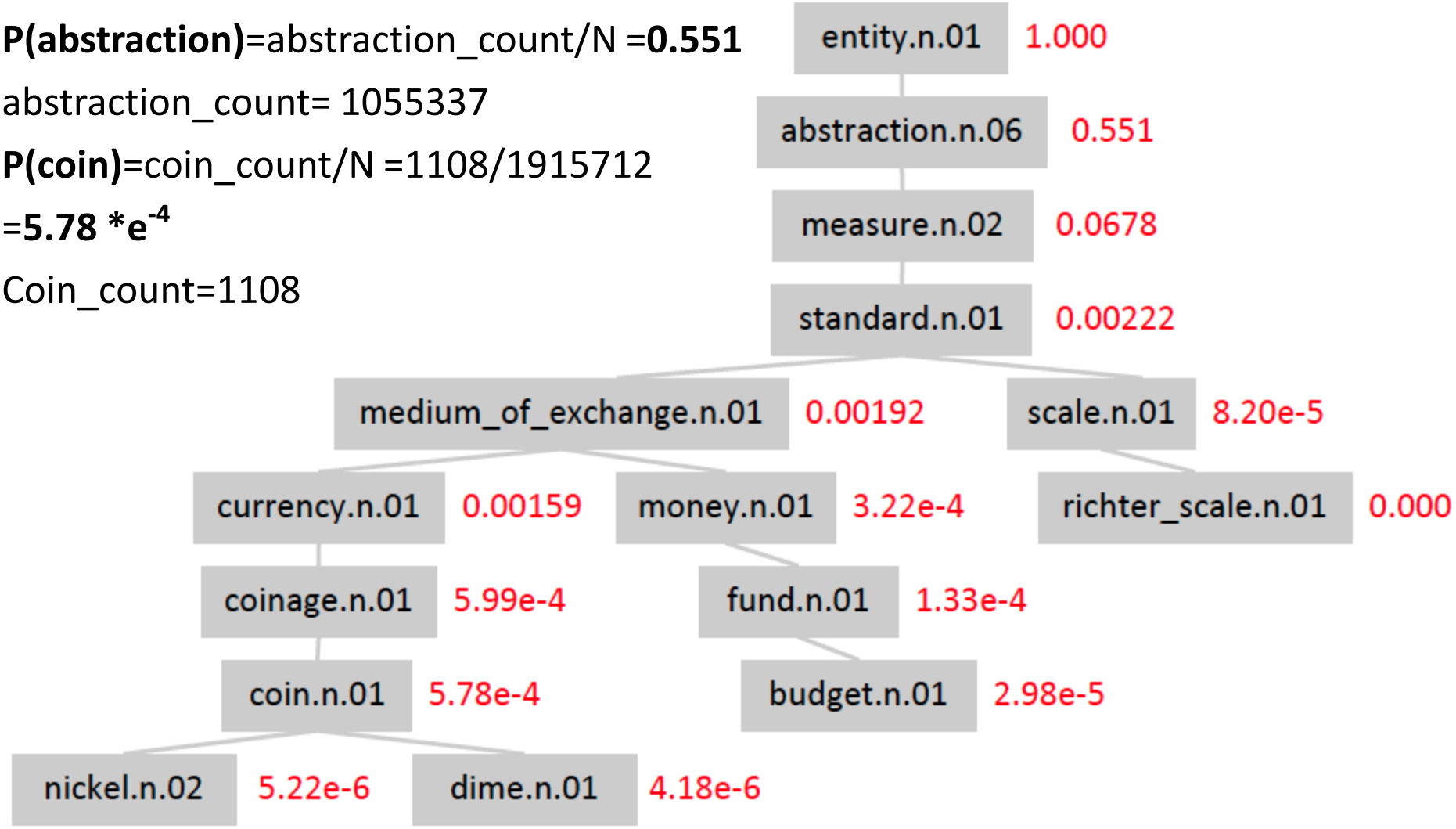
P(abstraction)=abstraction_count/N =**0.551**

abstraction_count= 1055337

P(coin)=coin_count/N =1108/1915712

=**5.78 *e^{-4}**

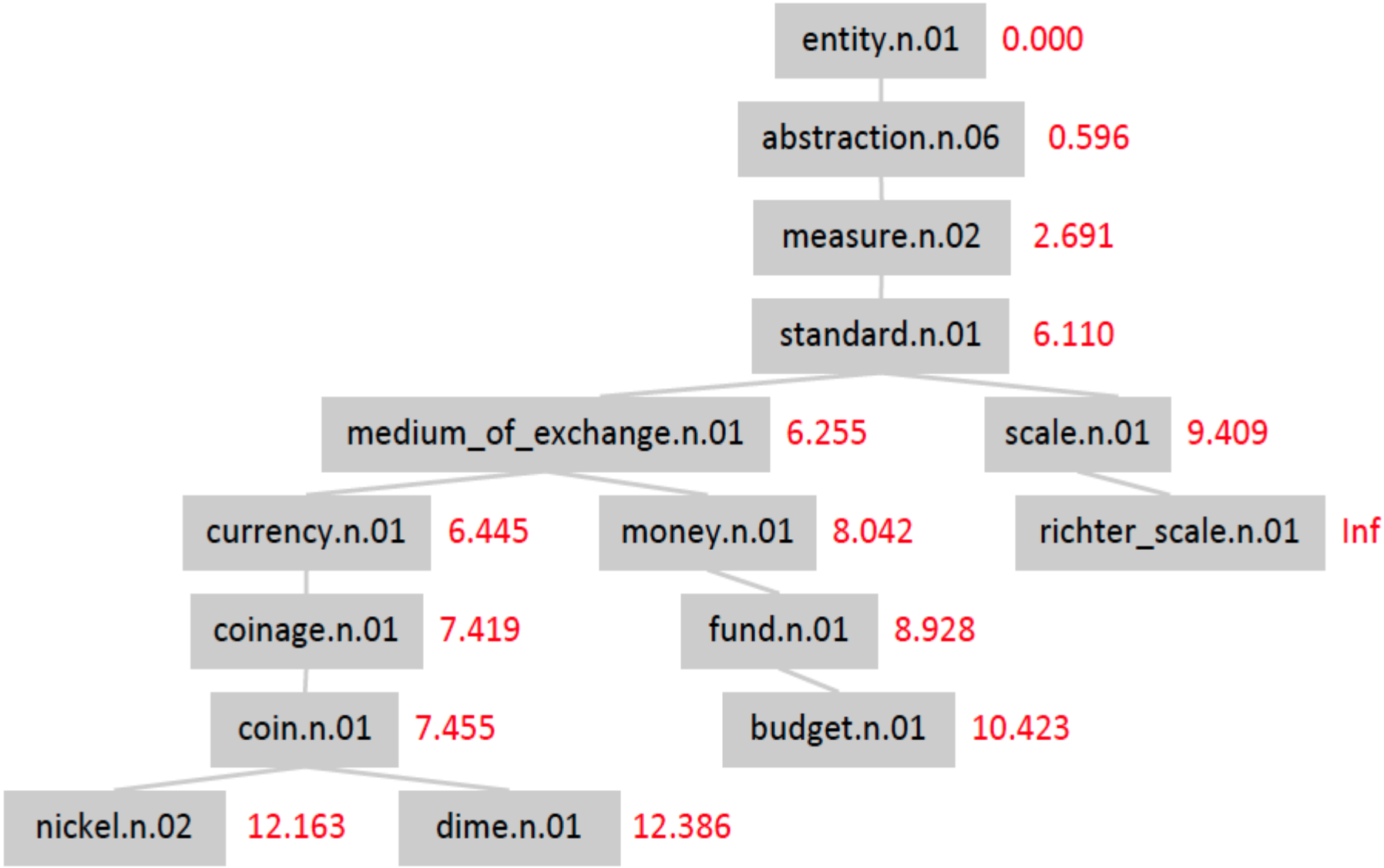
Coin_count=1108



Information Content

- The probabilities of concepts with very high value, do not have much information but if probability is low contains more information.
- Information content measures the specificity of a concept.
- **Information content: IC(c) = - log P(c)**
- **Lowest common subsumer** : LCS(c_1, c_2): the lowest node in the hierarchy that subsumes (is a hypernym of) both c1 and c2
- We are now ready to see how to use information content (IC) as a similarity metric.

Example : Information content

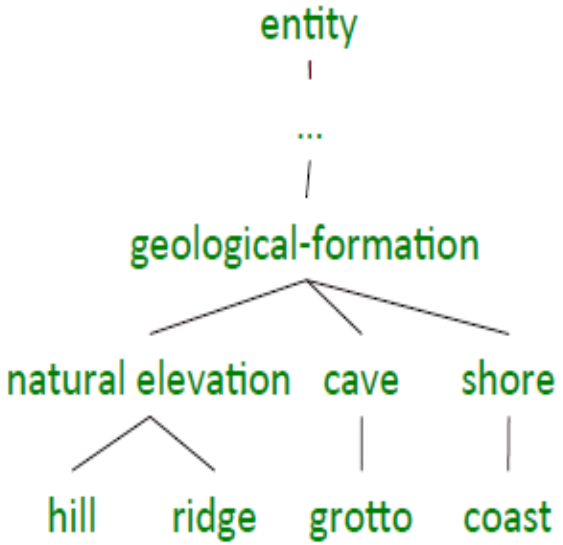


Concepts:

- **Subsumers:** the common ancestors of two synsets are called **subsumers** and the most specific ancestors among those is called as **lowest common subsume(lcs)**.
- **LCS(c1,c2) =** The most informative (lowest) node in the hierarchy subsuming both c1 and c2
- **Subsumers of hill and ridge** are:

{natural elevation, geological formation, …., entity}

- **LCS(hill, ridge) =** natural elevation



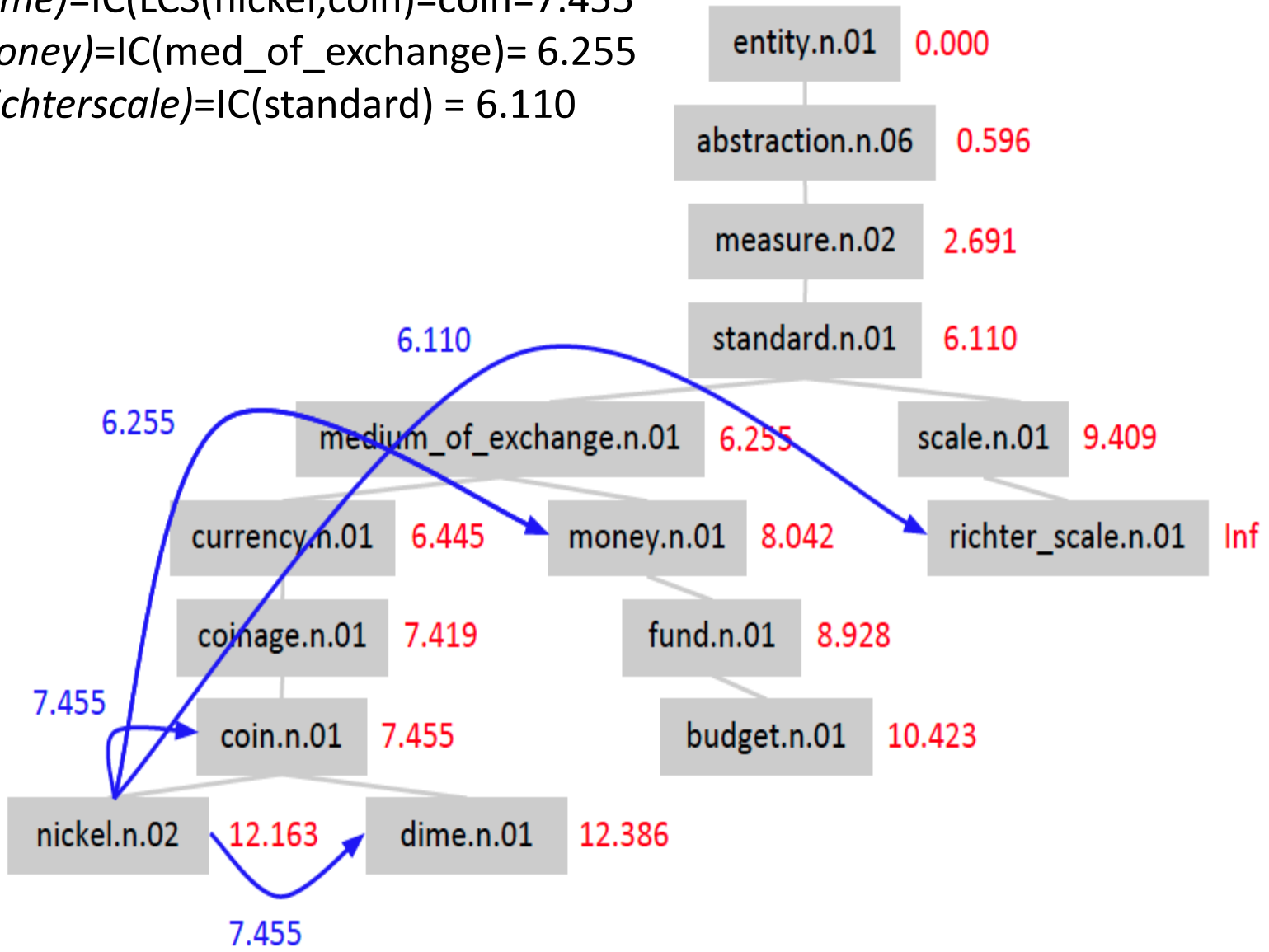
Resnik Similarity

- **Intuition:** how similar two words are depends on how much they have in common
- It measures the commonality by the information content of the lowest common subsumer

$$sim_{resnik}(c_1, c_2) = IC(LCS(c_1, c_2)) = -logP(LCS(c_1, c_2))$$

Example: Resnik similarity

sim_{resnik} (*nickel,coin)*=IC(LCS(nickel,coin)=IC(coin)=7.455
sim_{resnik} (*nickel,dime)*=IC(LCS(nickel,coin)=coin=7.455
sim_{resnik}(*nickel,money)*=IC(med_of_exchange)= 6.255
sim_{resnik} (*nickel,Richterscale)*=IC(standard) = 6.110



Problem with Resnik Similarity

- As we keep going up in hierarchy, the similarity decreases.
- As we keep doing down in hierarchy, the similarity increases. It is highest at single leaf node.
- sim_{resnik} (*coinage, money)* = sim_{resnik} (coinage*, budget)*
- This is because LCS for both is the same i.e., *medium_of_exchange .*
- *What is **captured** is how much information they share.*
- *What is **not captured** is how much information they don't share.*

Lin similarity

Proportion of shared information

- It's not just about commonalities - it's also about differences!
- **Resnik:** The more information content they share, the more similar they are
- **Lin**: The more information content they don't share, the less similar they are
- Not the absolute quantity of shared information but the proportion of shared information

$$sim_{Lin}(c_1, c_2) = \frac{2logP(LCS(c_1, c_2))}{logP(c_1) + logP(c_2)}$$

Lin similarity

$$sim_{Lin}(c_1, c_2) = \frac{2logP(LCS(c_1, c_2))}{logP(c_1) + logP(c_2)}$$

- The information content common to c_1 and c_2, normalized by their average information content.

sim_{Lin} (*coinage, money)* = 2*6.255/7.419+8.042

= 0.8091

sim_{Lin} (*coinage, budget)* = 2*6.255/7.419+10.423

= 0.7012

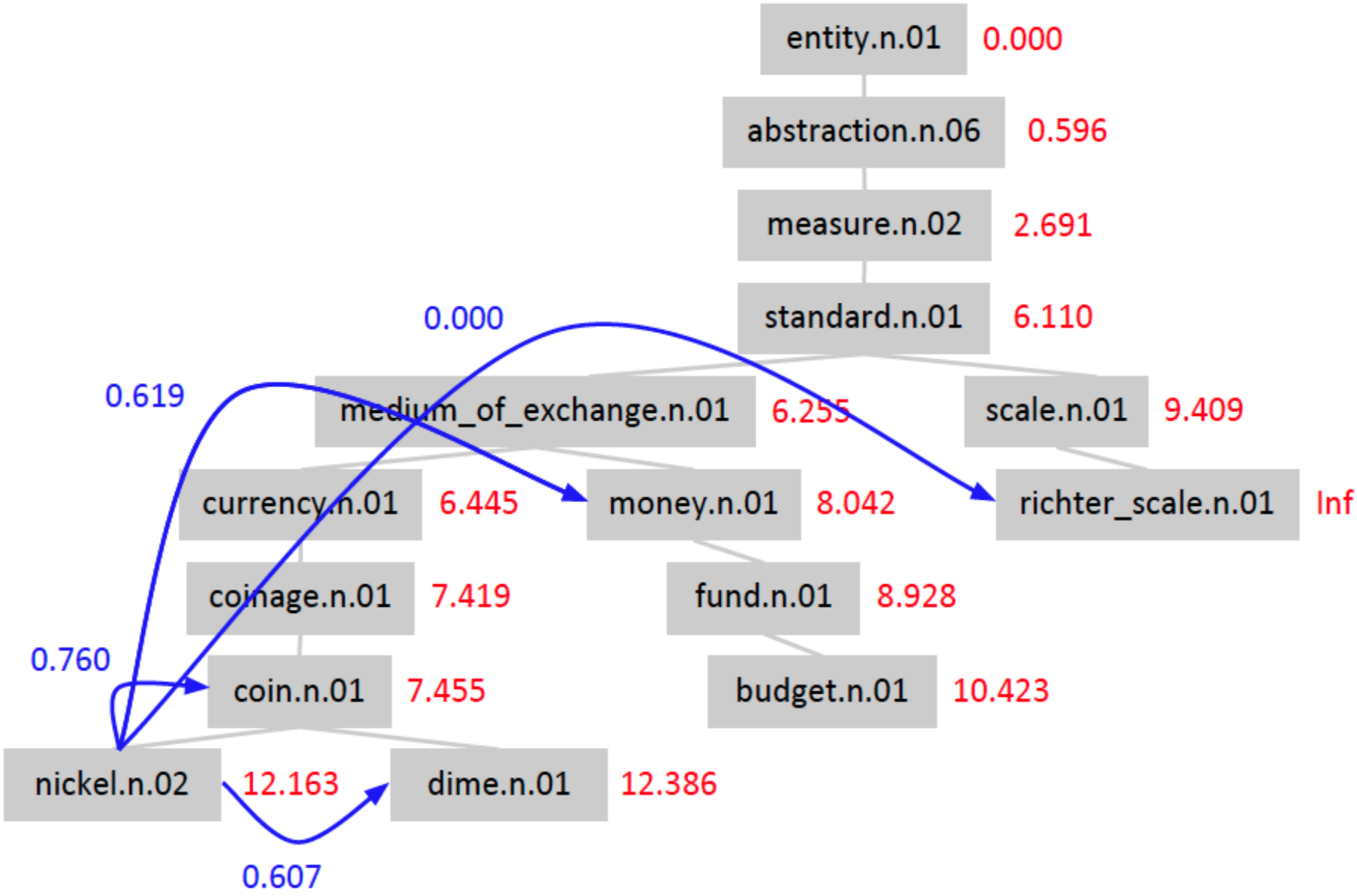
sim_{Lin} (*nickel, money)* = 2*6.255/12.163+8.042

= 0.6192

sim_{Lin} (*nickel, coin)* = 2*7.455/12.163+7.455

= 0.7600

Example: Lin similarity



Jiang-Conrath distance

JC similarity:

- We can use IC to assign lengths to graph edges:

$$dist_{JC}(c, hypernym(c)) = IC(c) - IC(hypernym(c))$$

$$\begin{aligned} dist_{JC}(c_1, c_2) &= dist_{JC}(c_1, LCS(c_1, c_2)) + dist_{JC}(c_2, LCS(c_1, c_2)) \\ &= IC(c_1) - IC(LCS(c_1, c_2)) + IC(c_2) - IC(LCS(c_1, c_2)) \\ &= IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2)) \end{aligned}$$

$$sim_{JC}(c_1, c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(LCS(c_1, c_2))}$$

Example: Jiang-Conrath distance

sim_{JC} (*nickel, coin)* = 1/12.163+7.455 –(2*7.455)

= 0.212

sim_{JC} (*nickel, money)* = 1/12.163+8.042 – (2*6.255)

= 0.13