I 202: Information Organization & Retrieval Fall 2025

Class 12: Semantic Similarity

Today's Outline

- Lexical Relations Review
- Taxonomic vs Thematic
- Semantic Similarity
- Card Sorting / Lumping vs Splitting
- Automating Semantic Similarity

DIFFERENT LEXICAL FORM

SAME LEXICAL FORM

hypernyms (superordinate)

fundamental measure

---

hypernyms (superordinate)

synonyms

time

polysemes

period (menstrual)

period

(amount of time)

sibling terms

temperature, mass, length

homographs

period (punctuation)

hyponyms (subordinate)

week, past, eve…

trial period, test period

hyponyms (subordinate)

RELATION: ANTONYMY

- Senses that are opposites with respect to only one feature
- Otherwise, they are very similar!

  dark/light *short/long* *fast/slow* *rise/fall*
  hot/cold *up/down* *in/out*

- More formally: antonyms can
  - *define a binary opposition or be at opposite ends of a scale*
    - `long/short, fast/slow`
  - *Be reversives:*
    - `rise/fall, up/down`

Slide from Jurafsky

CONNOTATION (SENTIMENT)

- Words have **affective** meanings
  - *Positive connotations (happy)*
  - *Negative connotations (sad)*
- Connotations can be subtle:
  - *Positive connotation: copy, replica, reproduction*
  - *Negative connotation: fake, knockoff, forgery*
- Evaluation (sentiment!)
  - *Positive evaluation (great, love)*
  - *Negative evaluation (terrible, hate)*

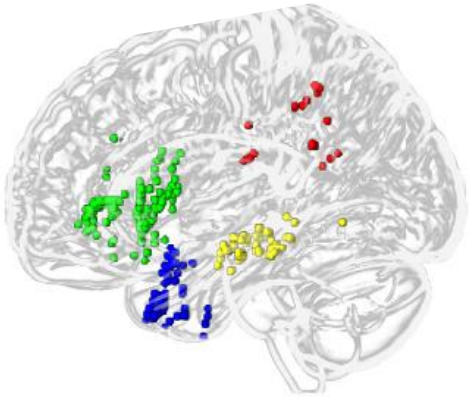
Slide from Jurafsky

CONNOTATION

- Some words seem to vary along 3 affective dimensions:
  - ***valence**: the pleasantness of the stimulus*
  - ***arousal**: the intensity of emotion provoked by the stimulus*
  - ***dominance**: the degree of control exerted by the stimulus*

	Word	Score		Word	Score
Valence	love	1.000		toxic	0.008
	happy	1.000		nightmare	0.005
Arousal	elated	0.960		mellow	0.069
	frenzy	0.965		napping	0.046
Dominance	powerful	0.991		weak	0.045
	leadership	0.983		empty	0.081

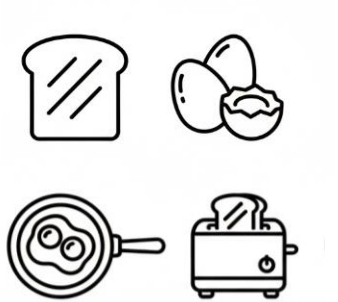
Osgood et al. 1957; Values from NRC VAD Lexicon (Mohammad 2018); Slide from Jurafsky



TAXONOMIC VS THEMATIC

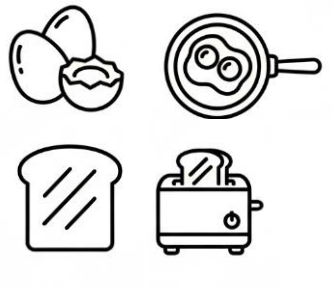
Taxonomic Vs Thematic Associations

- Taxonomic: Hyponym/Hypernym, IS-A
  - Eggs are a food, Toast is a food
  - Toaster is a cooking implement; so is a Pan



- Thematic: Concepts linked by some other relationship

  (used-for, made-from, etc.)
  - Eggs are cooked in a Pan
  - Bread is cooked in a Toaster



Icons by Gemini

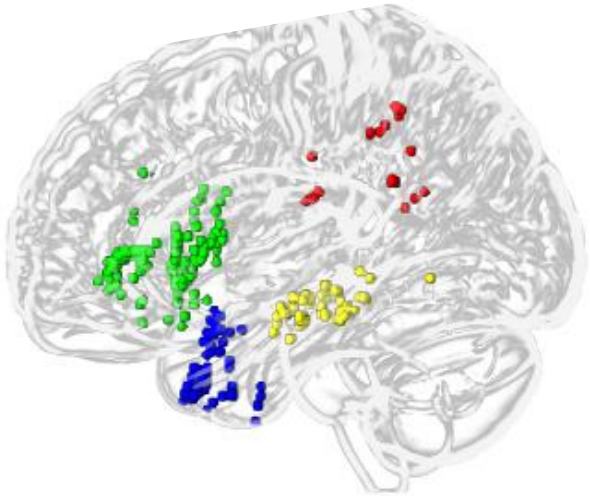
Taxonomic Vs Thematic Associations Evidence for Different Specialization in Regions in the Brain

Intracranial EEG readings suggest:

ATL specialized for taxonomic relations

IPL specialized for thematic relations

Close coordination is also suggested between the two regions.



Anterior Temporal Lobe (blue)
Inferior Parietal Lobule (red)

Thye et al, Intracranial EEG evidence of functional specialization for taxonomic and thematic relations, Cortex 140, 2021

Application:
Taxonomic vs Thematic Faceted Metadata

- Fine arts collection:
  - *Taxonomic: Weapons > Swords; Occupation > Soldier, etc*
  - *Thematic: Military Conflict (combines relevant subfacets)*
- Sporting goods vendor:
  - *Taxonomic: Shoes > Boots > Ski Boots; Outerwear > Jackets > Ski Jackets, etc*
  - *Thematic: Sport > Skiiing (combines relevant subfacets)*

Example: A Word-Association Study: Taxonomic Vs Thematic Associations

- Word pairs shown to crowd workers
- Rate the similarity from 1 -- 7

A Word-Association Study: Taxonomic Vs Thematic Associations

Taxonomic Instructions: Two words are similar if they look alike or belong to the same category.

- *For example, DOTS and STRIPES are similar (both are types of patterns or designs).*
- *However, SHIRT and STRIPES would not be similar.*

Thematic Instructions: Two words are connected or related if they occur in the same time or place; however, this does not mean they will share similar physical features.

- *For example, HELMET and MOTORCYCLE are related*
- *But CHRISTMAS-TREES and PALM-TREES are not related.*

Taxonomic Properties

High Taxonomic, Low Thematic

Word 1	Word 2
Breakfast	Dinner
Helmet	Crown
Salt	Sugar

Thematic Relations

Low Taxonomic, High Thematic

Word 1	Word 2
Helicopter	Pilot
Floss	Teeth
Pillow	Head

Landrigan & Mirman, Taxonomic and Thematic Relatedness Ratings for 659 Word Pairs, JOPD 4, 2016

Both Taxonomic and Thematic

High Taxonomic, High Thematic

Word 1	Word 2
Ring	Bracelet
Shingle	Brick
Tape	Staple

Ring, Bracelet:
Very close taxonomically

Thematic:
In this case, close taxonomically mirrors closeness thematically (similar usage)
(not the case for, e.g., tent & battery)

Neither Taxonomic Nor Thematic Similarity

Low Taxonomic, Low Thematic

Word 1	Word 2
Portrait	Report
Prisoner	Pupil
Bird	Lamb

Landrigan & Mirman, Taxonomic and Thematic Relatedness Ratings for 659 Word Pairs, JOPD 4, 2016

Let’s Do Some Card Sorting!
Groups of 3

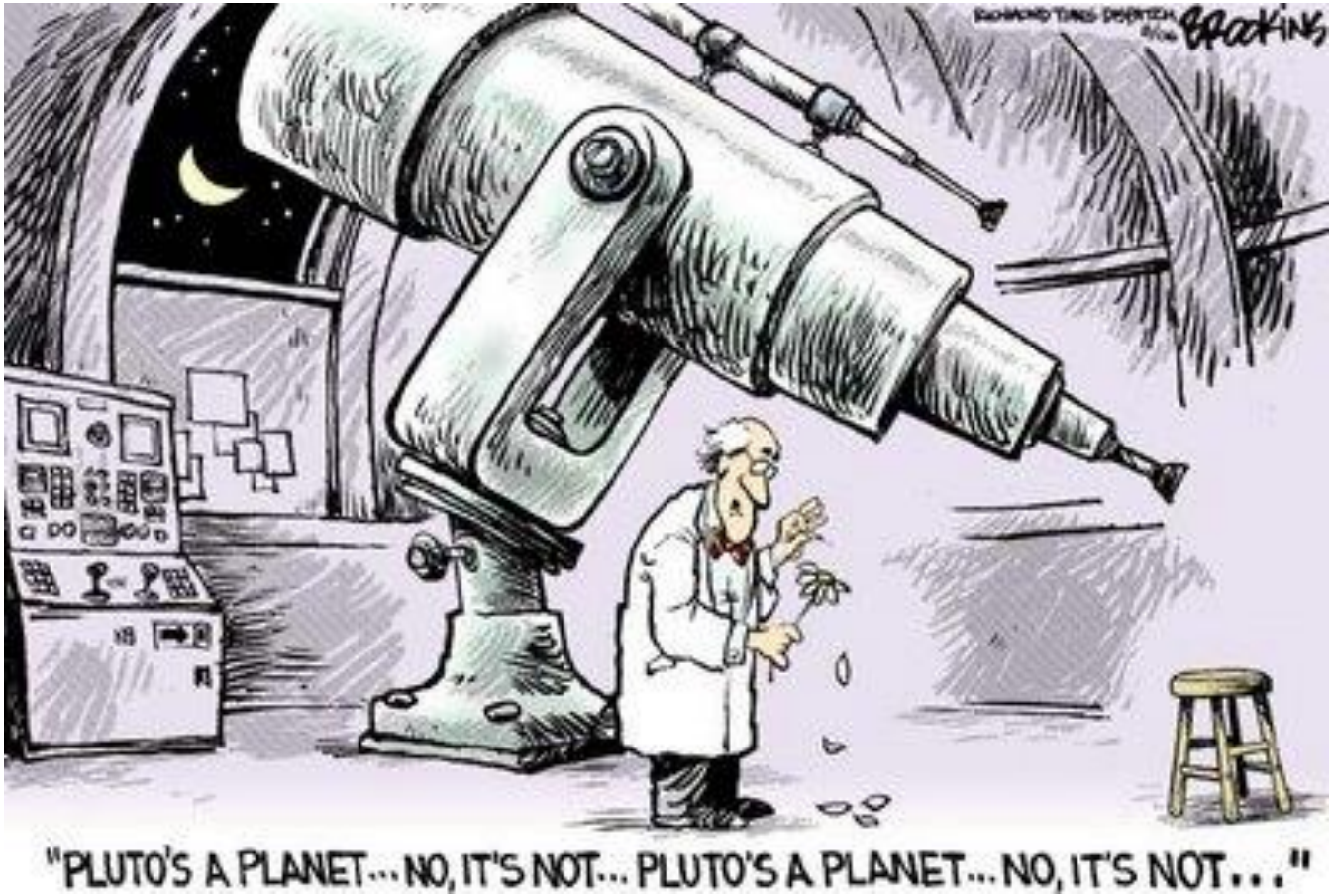
10 students use this link:

https://study.kardsort.com/cardsort2

10 students use this link:

https://tinyurl.com/3chr6ncr

Did you Lump or Split?



Gary Brookins/Richmond Times-Dispatch via NYTimes March 2015

LUMPERS AND SPLITTERS

"A **lumper** takes things that seem disparate and combines them because they have something similar. A **splitter** tends to take two things that are lumped together and separate them into smaller categories."

--Seth Maislin, American Society of Indexers, quoted by David Weinberger, *Everything Is Miscellaneous*, pg. 71, Times Books, 2007

We Constantly have to make Lump vs Split Decisions

How we organize our clothes

- *A big pile in the closet*
- *Old vs new*
- *Organized the Marie Kondo way*

- How we write up reports of a usability study?
  - *Lots of detail, broken out into tables?*
  - *Overall impressions, generalizations?*

The Miscellaneous Category

There is almost always an “other” or “miscellaneous” category

Our organizing systems do not handle these well

David Weinberger, *Everything Is Miscellaneous*, Times Books, 2007

Strategy: Taxonomic Grouping, Coarse-Grained (lumped)

shoes	sporting equipment	Apparel	Accessories
running shoes	mountain bikes	wet suits	sunscreen
hiking boots	canoes		coolers
water shoes	kayaks		
	tents		
	oars		

Strategy: Taxonomic, Fine-Grained (split)

Footwear	Bicycles	Accessories	Boats
running shoes	mountain bikes	sunscreen	kayaks
water shoes		coolers	canoes
hiking boots			

Apparel	Boating Equipment	Camping Equipment
wet suits	oars	tents

Strategy: Thematic Grouping Coarse Grained (Lumped)

City Sports	mountain sports	Accessories	water sports
running shoes	mountain bikes	sunscreen	kayaks
	tents		oars
	hiking boots		wet suits
	coolers		water shoes
			canoes

SEMANTIC SIMILARITY

What makes two words or concepts similar?

RELATION: **SIMILARITY**

- Also called "word association"
- Words can be related in many ways
  - `coffee, tea`: **similar**
  - `coffee, cup`: **related**, not similar

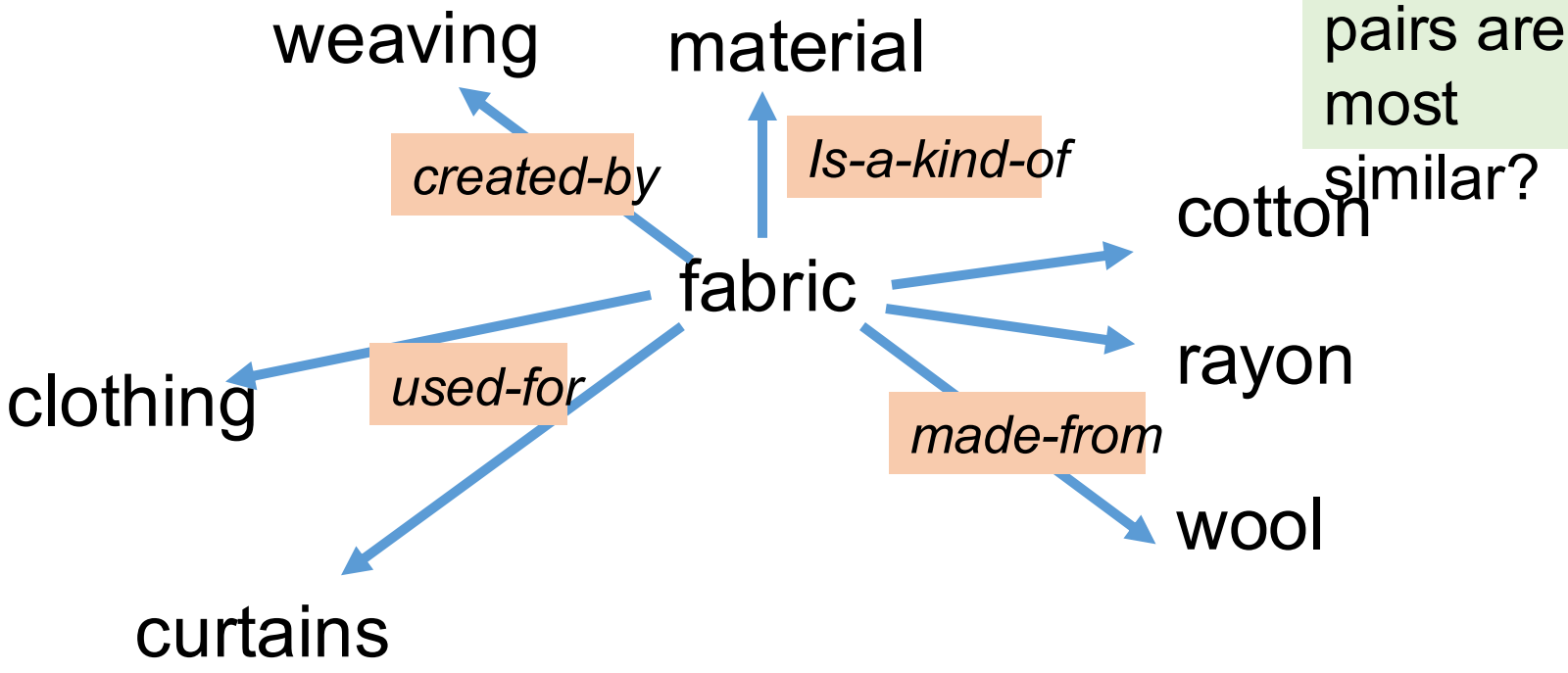
Results of Asking People How Similar 2 Words Are (from 0-10)

word1	word2	similarity
vanish	disappear	9.8
behave	obey	7.3
belief	impression	5.95
muscle	bone	3.65
modest	flexible	0.98
hole	agreement	0.3

SimLex-999 dataset (Hill et al., 2015)

Semantic Relations and Semantic Similarity

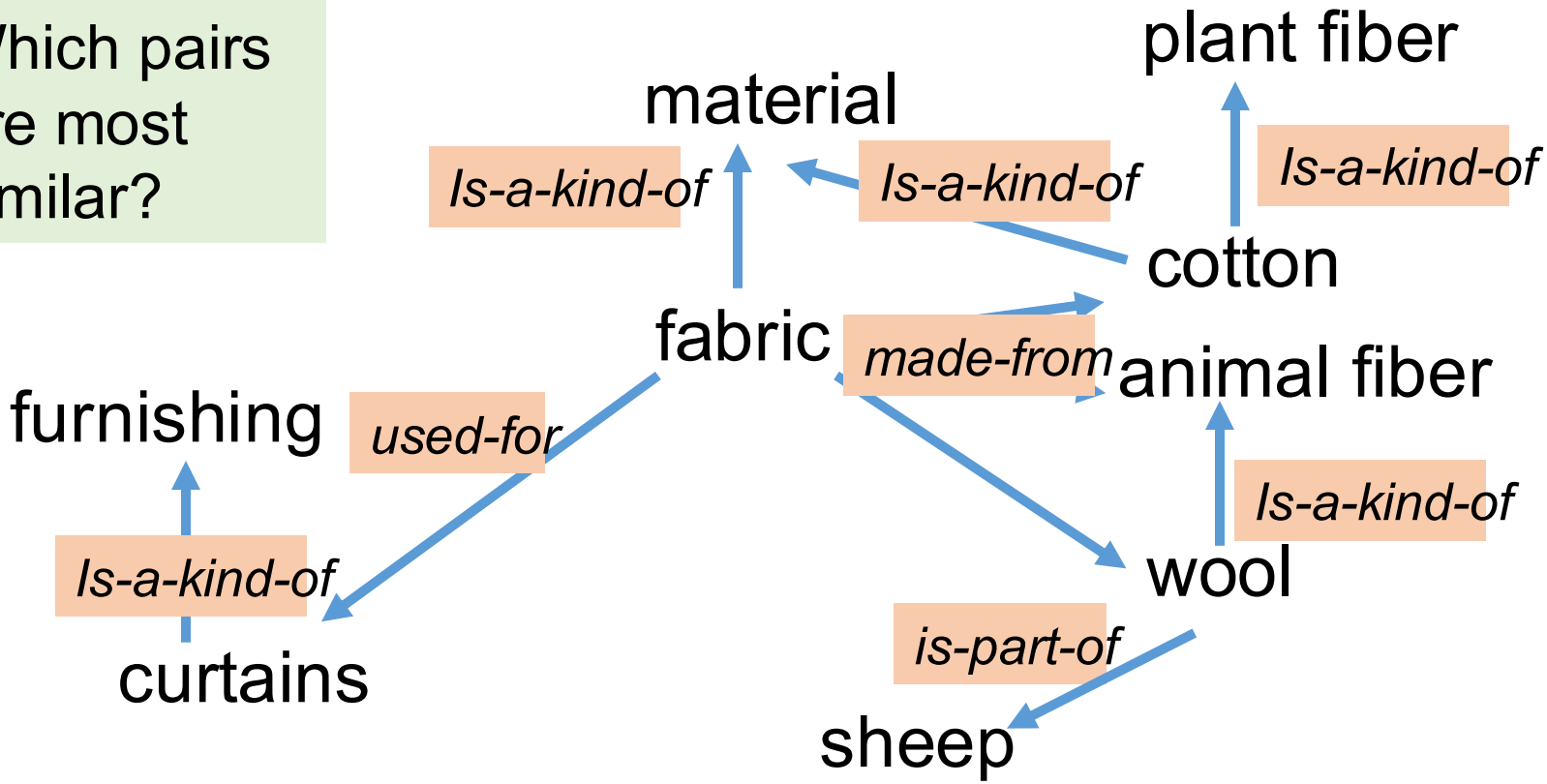
Different relation types may affect "similarity"



Which pairs are most similar?

Different relation types may affect "similarity"

Which pairs are most similar?



These Associations Influence Similarity Judgments
Which are the Most and Least Similar?

- Is *fabric* similar to *material*?
- Is *fabric* similar to *cotton*?
- Is *fabric* similar to *curtains*?
- Is *fabric* similar to *sheep*?
- Is *fabric* similar to *building*?
- Is *fabric* similar to *dream*?

Computing Similarity Values: Word Embeddings

- Represent words as vectors (arrays) of numbers
- Compare the values of the vectors to determine similarity
- The vectors try to capture the various types of relationships
- A closer similarity should correspond to having more shared features, weighted by category centrality.

Computing Word Similarity with Word Embeddings

```
compareWords(fabricPairs)
```

```
Similarity: fabric       material    : 0.5274
Similarity: fabric       cotton      : 0.7174
Similarity: fabric       curtain     : 0.5566
Similarity: fabric       building    : 0.2005
Similarity: fabric       dream       : 0.2140
```

Which relation types are scored as most similar?

Computed with Floret vector embeddings from Spacy (an extension of FastText, which is an extension of Word2Vec)

Code for Word Embedding-based Similarity

```
import spacy
```

```
nlp = spacy.load('en_core_web_lg')
```

```
def wordSim (word1, word2):
    vector1 = nlp(word1).vector
    vector2 = nlp(word2).vector

    # Reshape vectors for sklearn's cosine_similarity function (expects 2D arrays)
    v1_2d = vector1.reshape(1, -1)
    v2_2d = vector2.reshape(1, -1)

    # 2. Calculate the Cosine Similarity
    similarity = cosine_similarity(v1_2d, v2_2d)[0][0]

    # 3. Print the result
    print(f"Similarity: {word1:<12} {word2:<14}: {similarity:.4f}")
```

Computed with Floret vector embeddings from Spacy (an extension of FastText, which is an extension of Word2Vec)

Computing Word Similarity with Word Embeddings

```
compareWords(wordPairs)
```

```
Similarity: cat          dog          : 0.8017
Similarity: cat          siamese cat  : 0.8670
Similarity: cat          calico cat   : 0.8437
Similarity: cat          free cat     : 0.7873
Similarity: cat          lion         : 0.5265
Similarity: cat          feline       : 0.6990
Similarity: cat          scratch      : 0.3427
Similarity: cat          whiskers     : 0.3962
Similarity: cat          bark         : 0.3596
```

Which relation types are scored as most similar?

Computing Word Similarity with Word Embeddings

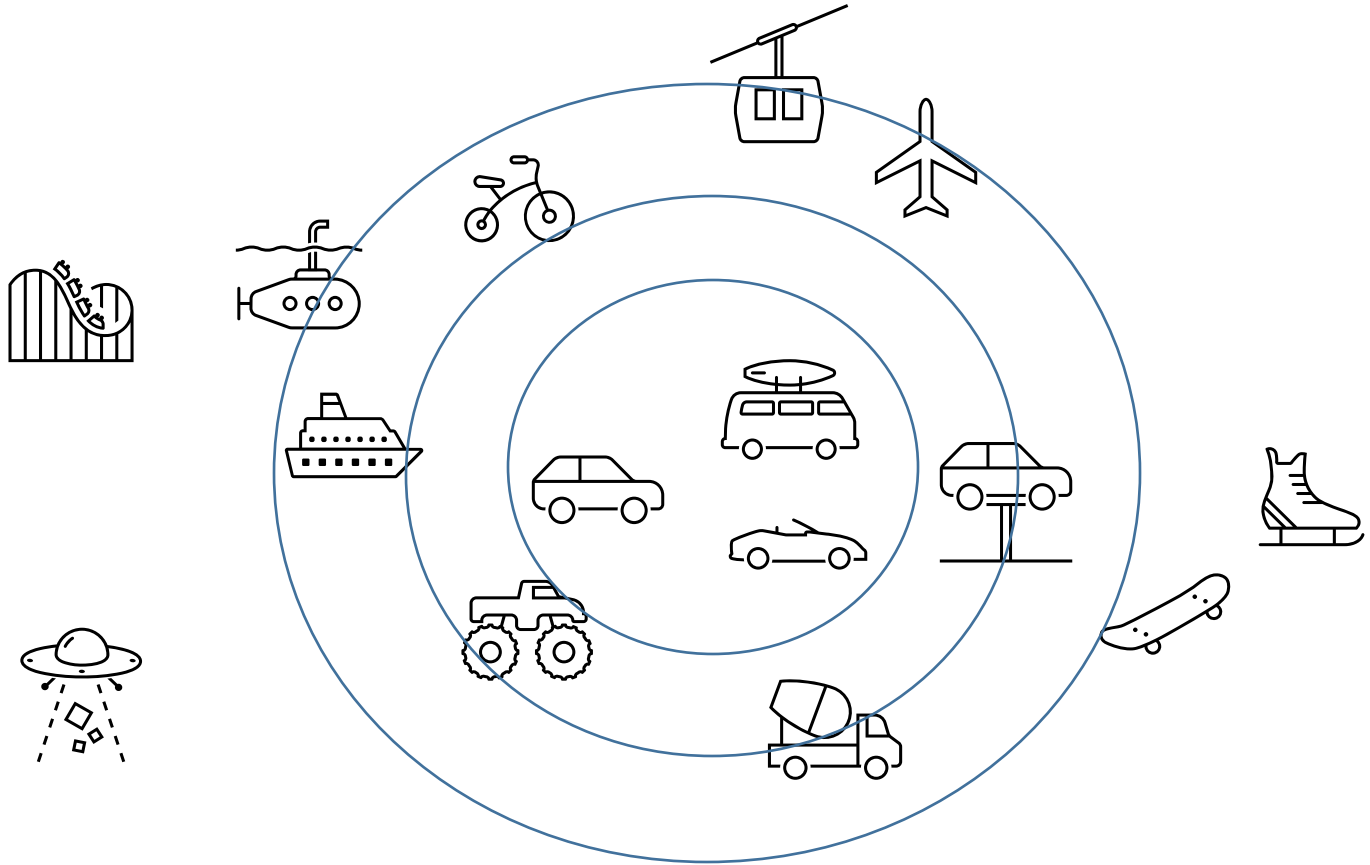
```
compareWords(autoPairs)
```

```
Similarity: car          auto        : 0.6832
Similarity: car          automobile  : 0.7536
Similarity: car          sedan       : 0.5304
Similarity: auto         sedan       : 0.3995
Similarity: auto         automobile  : 0.7419
Similarity: auto         automatic   : 0.5483
Similarity: car          automatic   : 0.3364
```

Example: Vehicle

More **Prototypical** is closer to the center

There is a **gradience** in category membership



Computing Word Similarity with Word Embeddings

```
compareWords(vehiclePairs)
```

```
Similarity: car          sports car    : 0.8354
Similarity: car          van           : 0.3788
Similarity: car          SUV           : 0.5542
Similarity: car          truck         : 0.7113
Similarity: car          monster car   : 0.8310
Similarity: car          cement mixer  : 0.2329
Similarity: car          cruise ship   : 0.3723
Similarity: car          repaired car  : 0.8164
Similarity: car          airplane      : 0.4787
Similarity: car          bicycle       : 0.4702
Similarity: car          submarine     : 0.1756
Similarity: car          funicular     : 0.1276
Similarity: car          roller coaster: 0.2762
Similarity: car          flying saucer : 0.3260
```

Computing Similarity Values: Sentence Embeddings

- Words in isolation can have many shades of meaning.
- The surrounding context of the word in a sentence clarifies the meaning.
- Sentence embeddings try to capture this meaning.
- Represents N words with N vectors (arrays) of numbers
- Average the N vectors to create one sentence embedding vector
- Compare the values of the 2 vectors to determine similarity as before

```
from sklearn.metrics.pairwise import cosine_similarity
from sentence_transformers import SentenceTransformer
```

```
def sentenceSim(s1, s2):
    embeddings = sentence_model.encode([s1, s2])
    embed1_2d = embeddings[0].reshape(1, -1)
    embed2_2d = embeddings[1].reshape(1, -1)
    sim_s1_s2 = cosine_similarity(embed1_2d, embed2_2d)[0][0]
    print(f"Similarity: {s1:<40} {s2:<40} {sim_s1_s2:.4f}")
```

Computing Similarity with Sentence Embeddings

```
compareSentences(treeSentences)
```

```
Similarity: There is some grass in a meadow    There is a tree in the meadow          0.8044
Similarity: There’s a tall patch of grass.     The root of the tree is in soil.       0.2971
Similarity: There's a tall patch of grass.     Let’s chat about pizza and cake!       0.0782
```

The similarity scores capture both lexical and conceptual similarity.

Computed with all-MiniLM-L6-v2, a small, distilled version of a larger Transformer model from Hugging Face.

Computing Similarity with Sentence Embeddings

```
compareSentences(catSentences)
```

```
Similarity: There is a cute cat                 There is a cute dog                 0.6255
Similarity: There is a cute cat                 There is a cute siamese cat         0.8047
Similarity: There is a cute cat                 There is a cute dream               0.4292
Similarity: Watch out for that scary cat        Watch out for that scary lion       0.7107
Similarity: A cat is a mammal                   A feline is a mammal                0.7193
Similarity: A cat might scratch you             A lion might scratch you            0.7050
Similarity: A cat has whiskers                  A lion has whiskers                 0.7077
Similarity: A cat might bark                    A lion might bark                   0.7134
Similarity: A cat is a feline                   A cat is a dog                      0.7890
```

Midterm Next Week

- Open Book and Notes
- No internet, no AI, no talking with others
- 3 Hours, but expected to take 2
- You choose the time that you start the 3 hours
  - *Oct 16 through 20th*
- Study guide at the start of next week

Next Week

- How embeddings are built
- Basics of automated classification
- Midterm Review