THE UNIVERSITY OF MELBOURNE

Comp90042 Workshop Week 6

Table of Content

1. Lexical Semantics

2. Distributional Semantics

Question 1

1. Give illustrative examples that show the difference between:

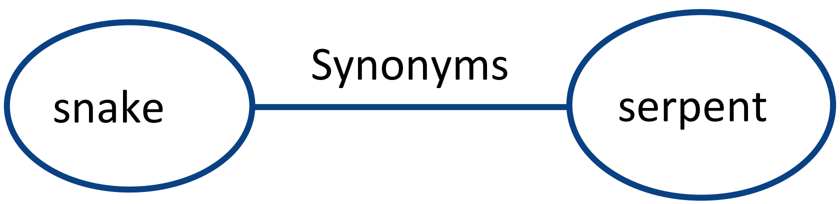
    (a) **Synonyms** and **hypernyms**

    (b) **Hyponyms** and **meronyms**

- The relationships between words meanings

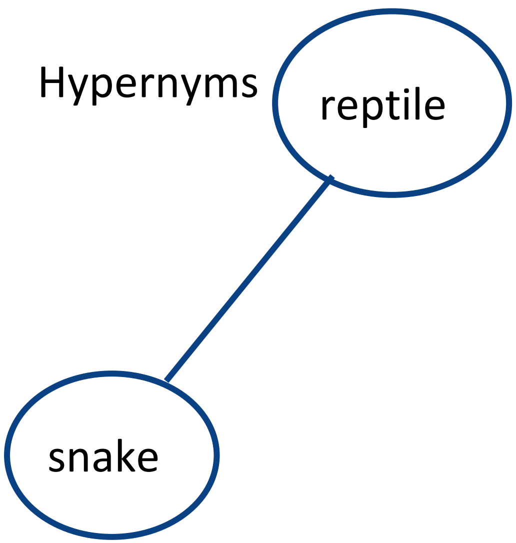
Synonyms

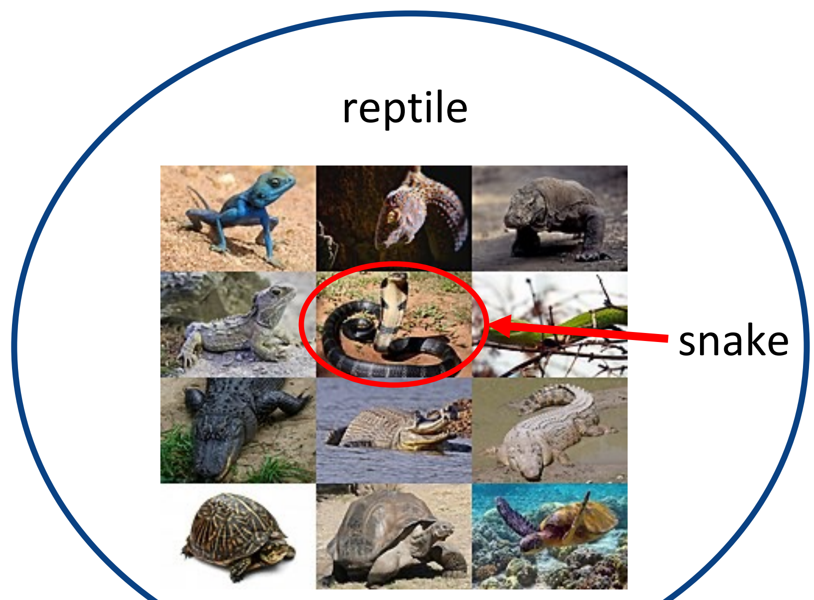
- **Synonyms:** words share (mostly) the same meanings
    - *snake and serpent*



Synonyms vs Hypernyms

- **Synonyms: words share (mostly) the same meanings**
  - ***snake and serpent***
- **Hypernyms:** One word is a hypernym of a second word when it is a more general instance ("higher up" in the hierarchy) of the latter
  - *reptile is the hypernym of snake (in its animal sense)*

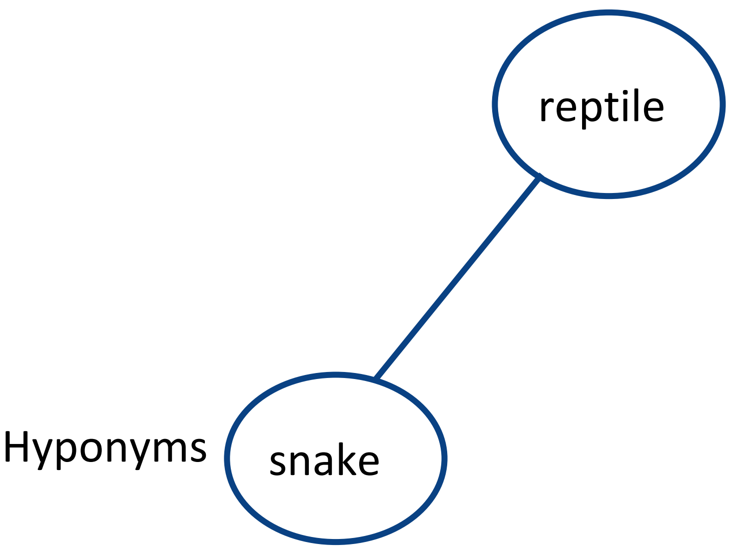


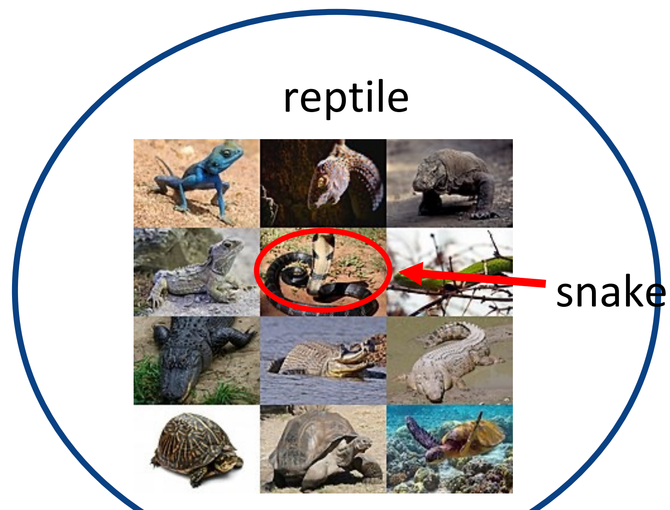


From https://en.wikipedia.org/wiki/Reptile

Hyponyms vs Hypernyms

- **Hypernyms:** One word is a hypernym of a second word when it is a more general instance ("higher up" in the hierarchy) of the latter
    - *reptile is the hypernym of snake (in its animal sense)*
- **Hyponyms :** One word is a hyponym of a second word when it is a more specific instance ("lower down" in the hierarchy) of the latter
    - *snake is the hyponym of reptile(in its animal sense)*



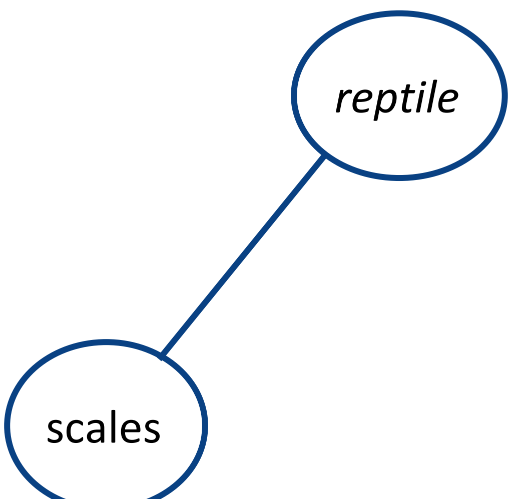


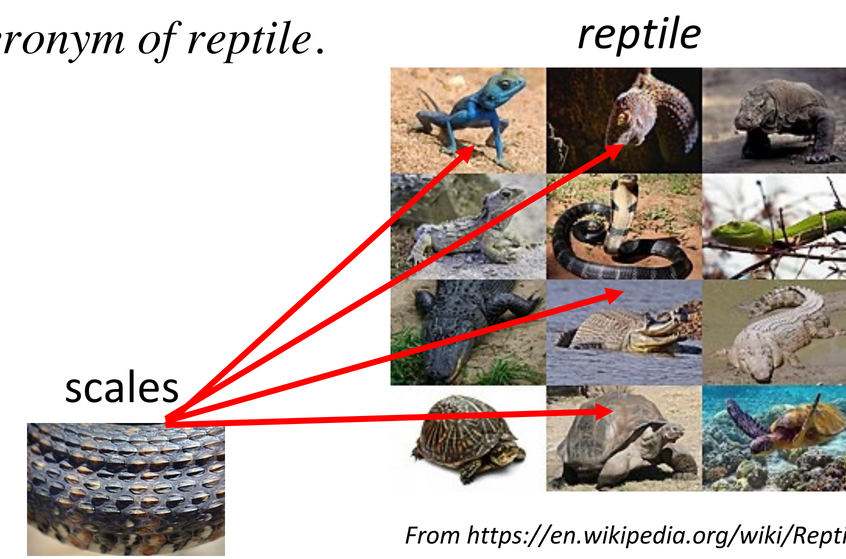
From https://en.wikipedia.org/wiki/Reptile

Hyponyms vs Meronyms

- **Hyponyms :** One word is a hyponym of a second word when it is a more specific instance ("lower down" in the hierarchy) of the latter
    - *snake is the hypernym of reptile(in its animal sense)*
- **Meronyms :** One word is a meronym of a second word when it is a part of the whole defined by the latter
    - *scales (the skin structure) is a meronym of reptile.*

Meronym



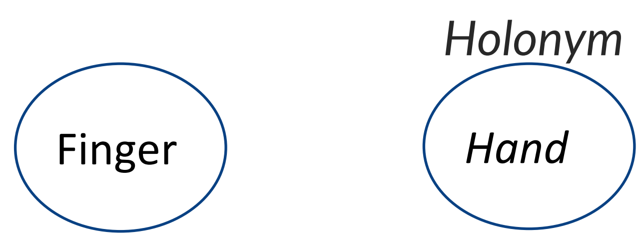


From https://en.wikipedia.org/wiki/Reptile

From https://en.wikipedia.org/wiki/Scale_(anatomy)

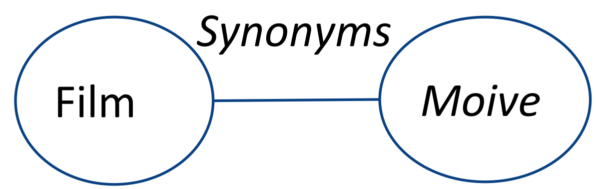
Exercise

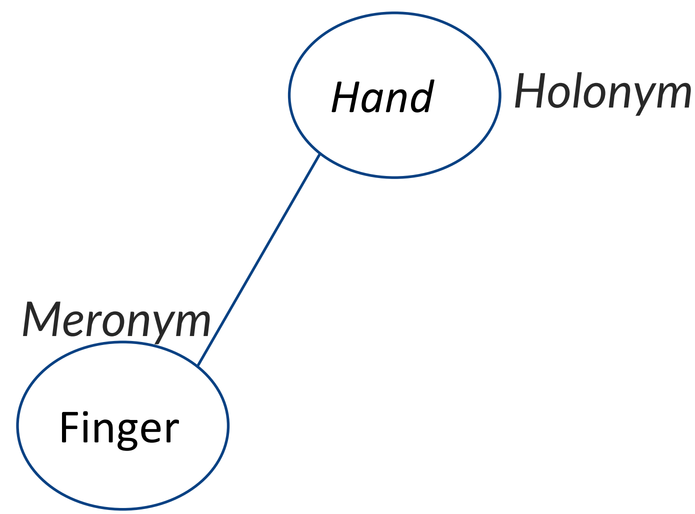
- **Movie and Film**
- **Hand and Finger**
- **Furniture and Table**

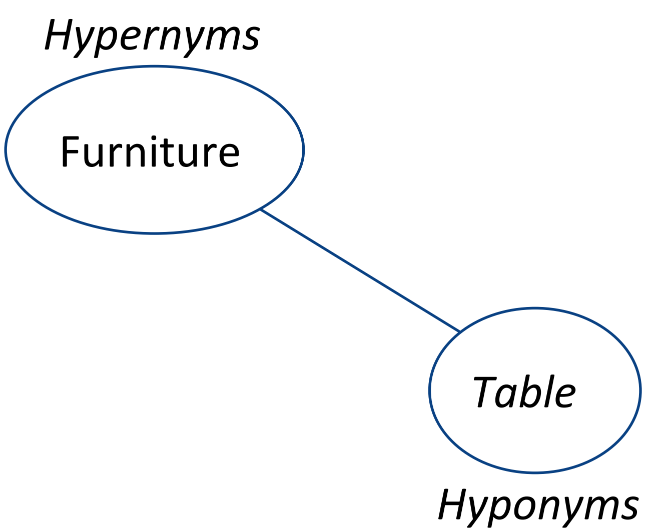


Exercise

- **Movie and Film**
- **Hand and Finger**
- **Furniture and Table**







Question 2

2. Using some Wordnet visualisation tool, for example, `http://wordnetweb.princeton.edu/perl/webwn` and the Wu & Palmer definition of **word similarity**, check whether the word *information* is more similar to the word *retrieval* or the word *science* (choose the sense which minimises the distance). Does this mesh with your intuition?

Wordnet

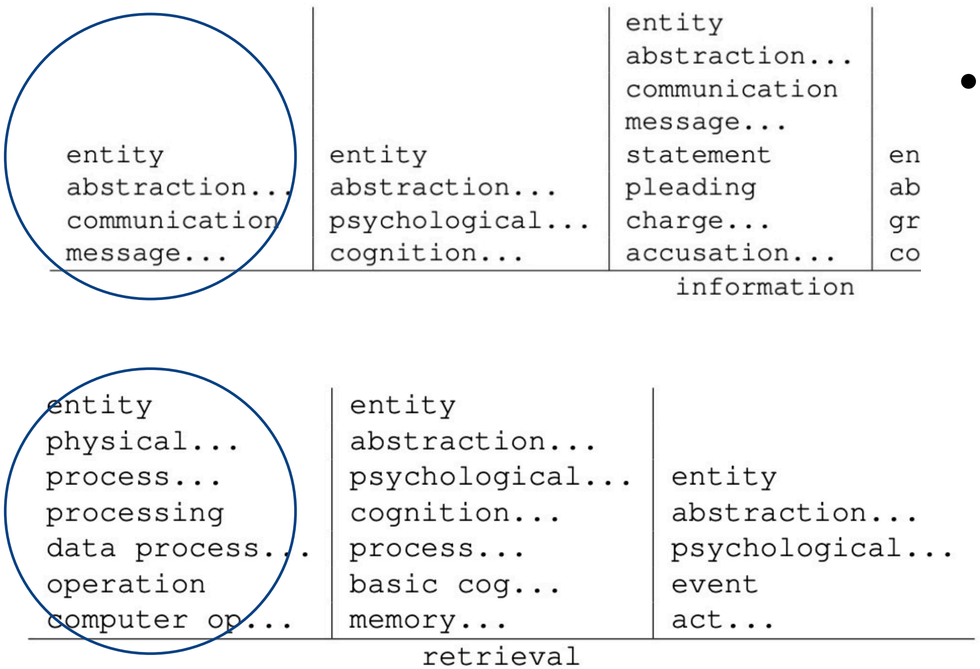
entity abstraction... communication message...	entity abstraction... psychological... cognition...	entity abstraction... communication message... statement pleading charge... accusation...	entity abstraction... group... collection...	entity abstraction... measure system of meas... information meas...

information

entity physical... process... processing data process... operation computer op...	entity abstraction... psychological... cognition... process... basic cog... memory...	entity abstraction... psychological... event act...

retrieval

Wordnet



- Wu & Palmer similarity

$$\text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

  - LCS: lowest common subsumer
  - Depth: path length from node to root

Wordnet

		entity	
		abstraction...	
		communication	
		message...	
entity	entity	statement	en
abstraction...	abstraction...	pleading	ab
communication	psychological...	charge...	gr
message...	cognition...	accusation...	co

information

entity	entity	
physical...	abstraction...	
process...	psychological...	entity
processing	cognition...	abstraction...
data process...	process...	psychological...
operation	basic cog...	event
computer op...	memory...	act...

retrieval

Choose the first meaning of two words

$$\mathrm{sim}(c_1, c_2) = \frac{2 \times \mathrm{depth}(\mathrm{LCS}(c_1, c_2))}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)}$$

Depth(information) =

Depth(retrieval) =

LCS(information, retrieval):

Depth(LCS) =

Wordnet

		entity	
		abstraction...	
		communication	
		message...	
entity	entity	statement	en
abstraction...	abstraction...	pleading	ab
communication	psychological...	charge...	gr
message...	cognition...	accusation...	co

information

entity	entity	
physical...	abstraction...	
process...	psychological...	entity
processing	cognition...	abstraction...
data process...	process...	psychological...
operation	basic cog...	event
computer op...	memory...	act...

retrieval

Choose the first meaning of two words

$$\text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

Depth(information) = 5

Depth(retrieval) = 8

LCS(information, retrieval):

Depth(LCS) =

Wordnet

		entity	
		abstraction...	
		communication	
		message...	
entity	entity	statement	en
abstraction...	abstraction...	pleading	ab
communication	psychological...	charge...	gr
message...	cognition...	accusation...	co

information

entity	entity	
physical...	abstraction...	
process...	psychological...	entity
processing	cognition...	abstraction...
data process...	process...	psychological...
operation	basic cog...	event
computer op...	memory...	act...

retrieval

Choose the first meaning of two words

$$\text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

Depth(information) = 5

Depth(retrieval) = 8

LCS(information, retrieval): entity

Depth(LCS) =

Wordnet

		entity	
		abstraction...	
		communication	
		message...	
entity	entity	statement	en
abstraction...	abstraction...	pleading	ab
communication	psychological...	charge...	gr
message...	cognition...	accusation...	co

information

entity	entity	
physical...	abstraction...	
process...	psychological...	entity
processing	cognition...	abstraction...
data process...	process...	psychological...
operation	basic cog...	event
computer op...	memory...	act...

retrieval

Choose the first meaning of two words

$$\text{sim}(c_1, c_2) = \frac{2 \times \text{depth}(\text{LCS}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}$$

Depth(information) = 5

Depth(retrieval) = 8

LCS(information, retrieval): entity

Depth(LCS) = Depth(entity) = 1

Wordnet

		entity	
		abstraction...	
		communication	
		message...	
entity	entity	statement	en
abstraction...	abstraction...	pleading	ab
communication	psychological...	charge...	gr
message...	cognition...	accusation...	co

information

entity	entity	
physical...	abstraction...	
process...	psychological...	entity
processing	cognition...	abstraction...
data process...	process...	psychological...
operation	basic cog...	event
computer op...	memory...	act...

retrieval

Choose the first meaning of two words

$$\mathrm{sim}(c_1, c_2) = \frac{2 \times \mathrm{depth}(\mathrm{LCS}(c_1, c_2))}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)}$$

Depth(information) = 5

Depth(retrieval) = 8

LCS(information, retrieval): entity

Depth(LCS) = Depth(entity) = 1

$$\begin{aligned} \mathrm{sim}(\mathrm{information}, \mathrm{retrieval}) &= \frac{2 \times 1}{5 + 8} \\ &= \frac{2}{13} \approx 0.154 \end{aligned}$$

Wordnet

entity abstraction... communication message...	entity abstraction... psychological... cognition...	entity abstraction... communication message... statement pleading charge... accusation...	en ab gr co

information

entity physical... process... processing data process... operation computer op...	entity abstraction... psychological... cognition... process... basic cog... memory...	entity abstraction... psychological... event act...

retrieval

Choose the first meaning of two words

$$\mathrm{sim}(c_1, c_2) = \frac{2 \times \mathrm{depth}(\mathrm{LCS}(c_1, c_2))}{\mathrm{depth}(c_1) + \mathrm{depth}(c_2)}$$

Depth(information) = 5

Depth(retrieval) = 8

LCS(information, retrieval): entity

Depth(LCS) = Depth(entity) = 1

$$\begin{aligned}\mathrm{sim}(\mathrm{information}, \mathrm{retrieval}) &= \frac{2 \times 1}{5 + 8} \\ &= \frac{2}{13} \approx 0.154\end{aligned}$$

Similarity

entity abstraction... communication message...	entity abstraction... psychological... cognition...	entity abstraction... communication message... statement pleading charge... accusation...	entity abstraction... group... collection...	entity abstraction... measure system of meas... information meas...

information

entity physical... process... processing data process... operation computer op...	entity abstraction... psychological... cognition... process... basic cog... memory...	entity abstraction... psychological... event act...

retrieval

		information				
		1	2	3	4	5
	1	0.154	0.154	0.118	0.154	0.143
retrieval	2	0.308	**0.615**	0.235	0.308	0.286
	3	0.364	0.545	0.267	0.364	0.333

Similarity

- Try to calculating the similarity of "information" and "science" yourself

Similarity

- Try to calculating the similarity of "information" and "science" yourself

- The maximum similarity is 0.727

- sim(information, science) > sim (information, retrieval)

  - Does this mesh with your intuition?

Question 3

3. What is **word sense disambiguation**?

Question 3

3. What is **word sense disambiguation**?

bank^1 : ...a *bank* can hold the investments in a custodial account ...
bank^2 : ...as agriculture burgeons on the east *bank*, the river ...

- Words can have multiple senses

Question 3

3. What is **word sense disambiguation**?

bank[1] : ...a *bank* can hold the investments in a custodial account ...
bank[2] : ...as agriculture burgeons on the east *bank*, the river ...

- Words can have multiple senses
- Word sense disambiguation
  - automatically determining which sense (usually, Wordnet synset) of a word is intended for a given token instance with a document.

he sat on the bank of the river and watched the currents

Question 3

3. What is **word sense disambiguation**?

bank^1 : ...a *bank* can hold the investments in a custodial account ...
bank^2 : ...as agriculture burgeons on the east *bank*, the river ...

- Words can have multiple senses

- Word sense disambiguation
  - automatically determining which sense (usually, Wordnet synset) of a word is intended for a given token instance with a document.

he sat on the bank of the river and watched the currents

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

Point-wise Mutual Information (PMI)

- represent how often two events co-occur

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- $p(x, y)$: joint distribution of x and y $= count(x, y)/\Sigma$
- $p(x)$: individual distribution of x. $= \Sigma x \,/\Sigma$
- $p(y)$ individual distribution of y $= \Sigma y \,/\Sigma$

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- Σ (Total number of instance)
  $=$
- $P(world) =$
- $P(cup) =$
- $P(world, cup) =$

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- Σ (Total number of instance)

  $= 55 + 225 + 315 + 1405 = 2000$
- $P(world) =$
- $P(cup) =$
- $P(world, cup) =$

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

$$PMI(x,y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

- Σ (Total number of instance)
  $= 55 + 225 + 315 + 1405 = 2000$
- $P(world) = (55 + 225) / 2000 = 0.14$
- $P(cup) =$
- $P(world, cup) =$

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- Σ (Total number of instance)
  $= 55 + 225 + 315 + 1405 = 2000$
- $P(world) = (55 + 225) / 2000 = 0.14$
- $P(cup) = (55 + 315) / 2000 = 0.185$
- $P(world, cup) = 55 / 2000 = 0.0275$

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- Σ (Total number of instance) $= 55 + 225 + 315 + 1405 = 2000$
- $P(world) = (55 + 225) / 2000 = 0.14$
- $P(cup) = (55 + 315) / 2000 = 0.185$
- $P(world, cup) = 55 / 2000 = 0.0275$
- $PMI\ (world, cup) =$

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

$$PMI(x, y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

- Σ (Total number of instance)
  $= 55 + 225 + 315 + 1405 = 2000$
- $P(world) = (55 + 225) / 2000 = 0.14$
- $P(cup) = (55 + 315) / 2000 = 0.185$
- $P(world, cup) = 55 / 2000 = 0.0275$
- $PMI\ (world, cup) = \log_2 \frac{p(world,cup)}{p(world)*p(cup)}$
  $= \log_2 \frac{0.0275}{0.14*0.185}$
  ≈ 0.0865

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

Question 4

4. For the following term co-occurrence matrix (suitably interpreted):

	cup	not(cup)
world	55	225
not(world)	315	1405

(a) Find the Point-wise Mutual Information (PMI) between these two terms in this collection.

(b) What does the value from (a) tell us about **distributional similarity**?

- $PMI(w, c) \approx 0.0865$
- Distributional similarity
  - slightly positive
  - occur together slightly more commonly than would occur purely by chance.
  - World Cup!

Question 5

5. In the `09-distributional-semantics` iPython notebook, a term–document matrix is built to learn word vectors.

   (a) What is the Singular Value Decomposition (SVD) method used for here? Why is this helpful?

Singular Value Decomposition

$$A = U\Sigma V^T$$

$$\begin{array}{cc} & |D| \\ |V| & \begin{bmatrix} 0 & 1 & 0 & & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \end{array}$$

A
(term-document matrix)

U
(new term matrix)

Σ
(singular values)

V^T
(new document matrix)

$$\begin{array}{cc} & m \\ |V| & \begin{bmatrix} 2.2 & 0.3 & & 8.7 \\ 5.5 & -2.8 & \cdots & 0.1 \\ -1.3 & 3.7 & & 3.5 \\ \vdots & & \ddots & \vdots \\ 2.9 & -2.1 & \cdots & -1.9 \end{bmatrix} \end{array} \quad \begin{array}{cc} & m \\ m & \begin{bmatrix} 9.1 & 0 & 0 & & 0 \\ 0 & 4.4 & 0 & \cdots & 0 \\ 0 & 0 & 2.3 & & 0 \\ & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0.1 \end{bmatrix} \end{array} \quad \begin{array}{cc} & |D| \\ m & \begin{bmatrix} -0.2 & 4.0 & & -1.3 \\ -4.1 & 0.6 & \cdots & -0.2 \\ 2.6 & 6.1 & & 1.4 \\ \vdots & & \ddots & \vdots \\ -1.9 & -1.8 & \cdots & 0.3 \end{bmatrix} \end{array}$$

$m = Rank(A)$

12

Question 5

5. In the `09-distributional-semantics` iPython notebook, a term–document matrix is built to learn word vectors.

   (a) What is the Singular Value Decomposition (SVD) method used for here? Why is this helpful?

- throw away the less important characteristics
- identify the most important characteristics of word

Question 5

5. In the `09-distributional-semantics` iPython notebook, a term–document matrix is built to learn word vectors.

   (a) What is the Singular Value Decomposition (SVD) method used for here? Why is this helpful?

- throw away the less important characteristics
- identify the most important characteristics of word
- smaller representation of the words (dense matrix)
- Save time and storage

Question 6

6. What is a **word embedding** and how does it relate to **distributional similarity**?
   (a) What is the difference between a **skip-gram** model and a **CBOW** model?
   (b) How are the above models trained?

- Word embedding:

Question 6

6. What is a **word embedding** and how does it relate to **distributional similarity**?
   (a) What is the difference between a **skip-gram** model and a **CBOW** model?
   (b) How are the above models trained?

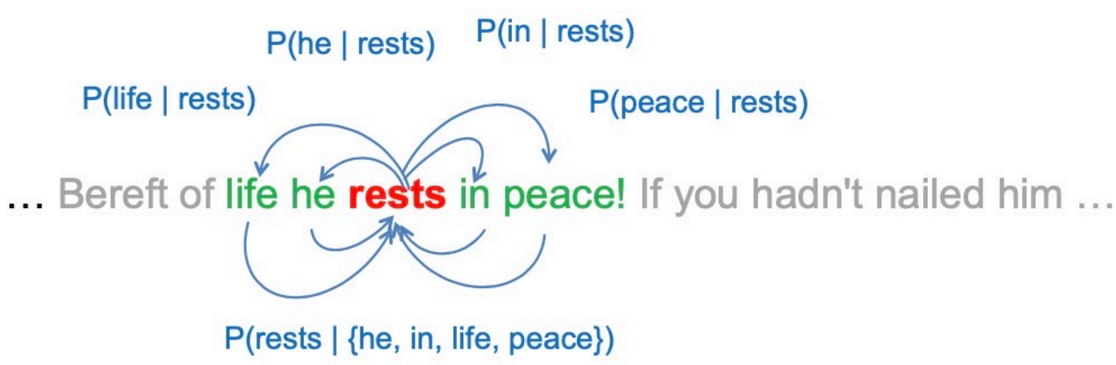
- Word embedding:
  - Representation of words into a low dimensionality vector space
  - Capture semantic and syntactic relationship between words

Question 6

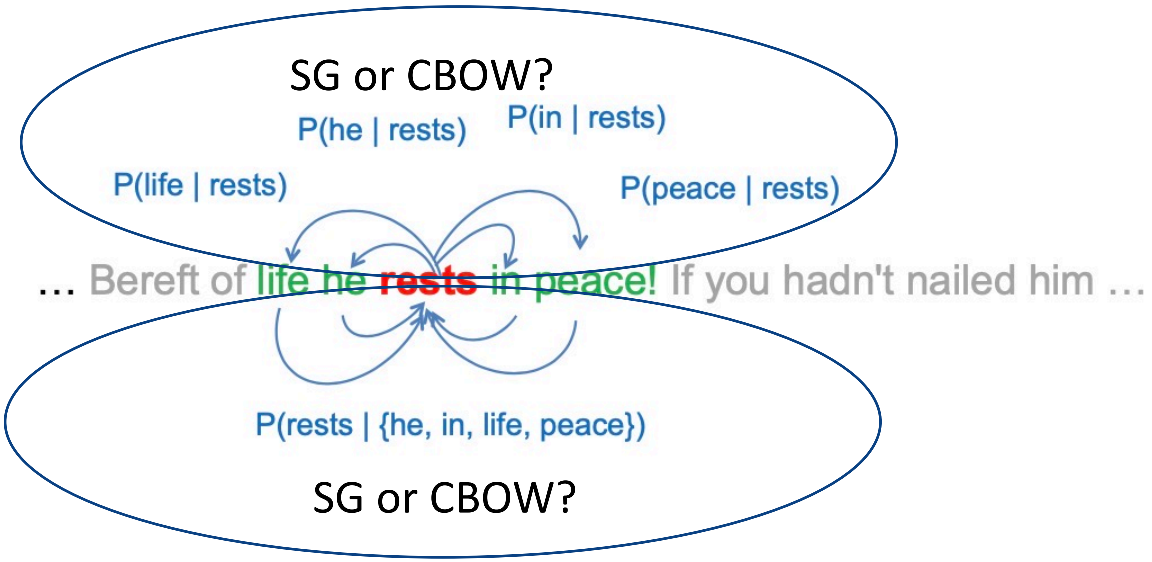
6. What is a **word embedding** and how does it relate to **distributional similarity**?
   (a) What is the difference between a **skip-gram** model and a **CBOW** model?
   (b) How are the above models trained?

- Word embedding:
  - Representation of words into a low dimensionality vector space
  - Capture semantic and syntactic relationship between words
  - broadly the same as what we expect in distributional similarity

SG and CBOW

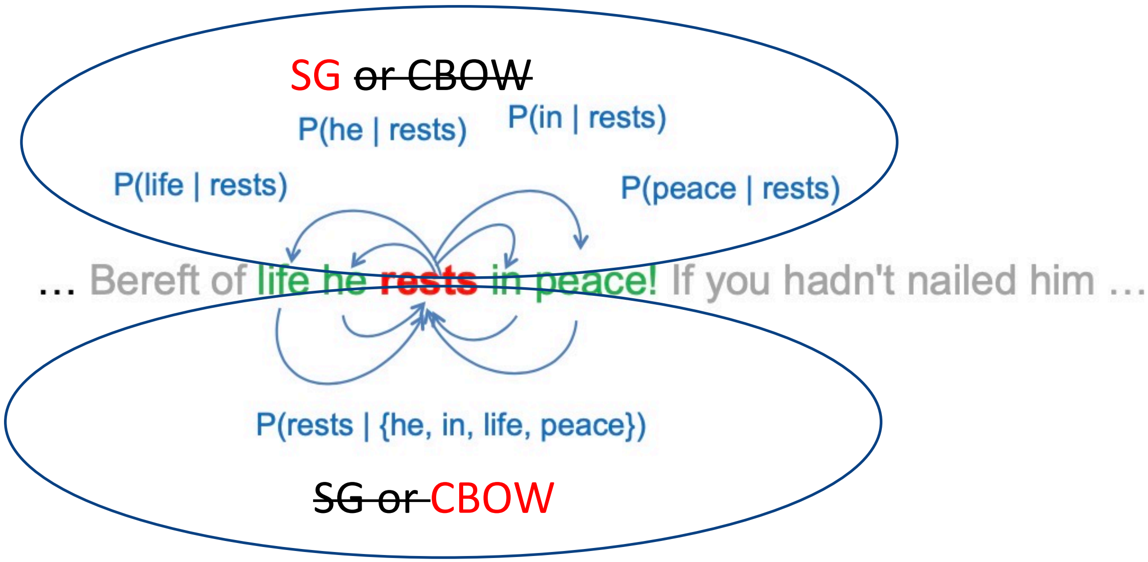

P(he | rests)
P(in | rests)
P(life | rests)
P(peace | rests)
… Bereft of life he rests in peace! If you hadn't nailed him …
P(rests | {he, in, life, peace})

SG and CBOW


SG or CBOW?
P(he | rests)
P(in | rests)
P(life | rests)
P(peace | rests)
… Bereft of life he rests in peace! If you hadn't nailed him …
P(rests | {he, in, life, peace})
SG or CBOW?

SG and CBOW

skip–gram models analyse the probability of the context words given the target word;



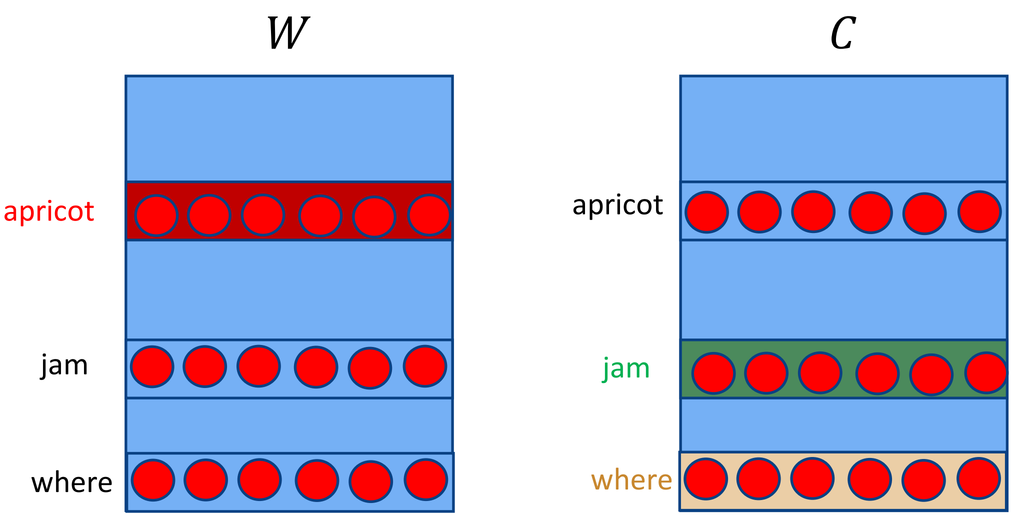
CBOW models analyse the probability of the target word given the context words.

Training

- Example of SG

```
... lemon,  a [tablespoon of apricot jam,     a] pinch ...
               c1         c2      w      c3      c4
```

windows =2



Target word: apricot

Context words: tablespoon, of, jam, a

noise words: where, seven, if (random sample from V)

Training

- Example of SG

```
... lemon,  a [tablespoon of apricot jam,     a] pinch ...
                c1          c2    w     c3      c4
```

positive examples +

w	c_{pos}
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

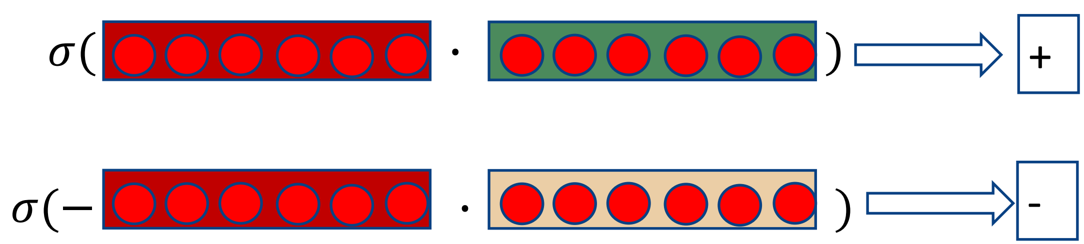
w	c_{neg}	w	c_{neg}
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

$$L_{CE} = -\log\left[P(+|w,c_{pos})\prod_{i=1}^{k}P(-|w,c_{neg_i})\right]$$

$\sigma: sigmoid(logistic)\ function$

Positive

Negative



Training

- Loss function

$$
\begin{aligned}
L_{CE} &= -\log\left[P(+|w,c_{pos})\prod_{i=1}^{k}P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos})+\sum_{i=1}^{k}\log P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos})+\sum_{i=1}^{k}\log\left(1-P(+|w,c_{neg_i})\right)\right] \\
&= -\left[\log\sigma(c_{pos}\cdot w)+\sum_{i=1}^{k}\log\sigma(-c_{neg_i}\cdot w)\right]
\end{aligned}
$$