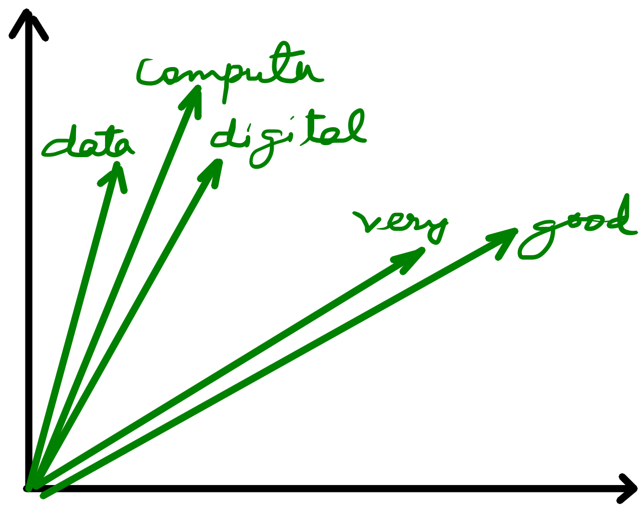
27-04-2021

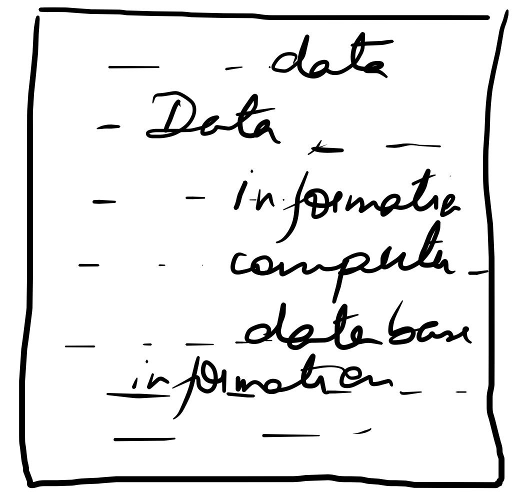
NLPIE

Word vectors or embeddings as vector of content values.



A word is defined by its surrounding words.

Co-occurance Matrix : A matrix in which each value f_{ij} represents the number of times word w_i is in content of c_j.



content words (content)

target words (words)

	data	stored	c_i memory	comp	system	
data	0	2	2	1	0	...
stored	2	0	2	2	0	...
w_i memory	1 2	1 2	f_{ij}	1 1	1 1	
computer						
system						

co-occurrence matrix with window size ±2 of size |V|

- we define a window of contents words.
- each content words is based on some window size.
- e.g. "Data is stored in memory of a computer system".

① Data stored memory computer system

$\underbrace{\text{Data stored}}$ memory $\underbrace{\text{computer system}}$

↑ target word

content words

② computer stores data in memory

computer stores data memory

outside the window of ±2, not considered as content

Vector Semantics

Pointwise Mutual Information (PMI): The measure how often two events x and y occur together, compared with if they occur independently

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Joint prop

Independent Prob

PMI between a terget word w and context word c may be defined as $\text{PMI}(w,c) = \log_2 \frac{P(w,c)}{P(w)P(c)}$

Co-occuring things less often leads to negative PMIvalues. However, we are concerned with co-occuring things vey often. Therefore, it is common to use Positive PMI,

$$\text{PPMI}(w,c) = \max(\log_2 \frac{P(w,c)}{P(w)P(c)}, 0)$$

Given a co-occurance metrix with W rows of words, C columns of contexts, and f_{ij} is the number of times word w_i occurs in context c_j. Then PPMI matrix may be defined with PPMI_{ij}

$$\text{PPMI}_{ij} = \max(\log_2 \frac{p_{ij}}{p_{i*}p_{*j}}, 0)$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

Co-occurrence counts for four words in 5 contexts in the Wikipedia corpus

	computer	data	result	pie	sugar	count(w)
cherry	2	8	9	442	25	486
strawberry	0	0	1	60	19	80
digital	1670	1683	85	5	4	3447
information	3325	3982	378	5	13	7703
count(context)	4997	5673	473	512	61	11716

$$P(\text{w=information,c=data}) = \frac{3982}{11716} = .3399$$

$$P(\text{w=information}) = \frac{7703}{11716} = .6575$$

$$P(\text{c=data}) = \frac{5673}{11716} = .4842$$

$$\text{ppmi(information,data)} = \log 2(.3399/(.6575 * .4842)) = .0944$$

$$p_{ij} = \frac{f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{i*} = \frac{\sum_{j=1}^{C} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}} \quad p_{*j} = \frac{\sum_{i=1}^{W} f_{ij}}{\sum_{i=1}^{W}\sum_{j=1}^{C} f_{ij}}$$

	p(w,context)					p(w)
	computer	**data**	**result**	**pie**	**sugar**	**p(w)**
cherry	0.0002	0.0007	0.0008	0.0377	0.0021	0.0415
strawberry	0.0000	0.0000	0.0001	0.0051	0.0016	0.0068
digital	0.1425	0.1436	0.0073	0.0004	0.0003	0.2942
information	0.2838	0.3399	0.0323	0.0004	0.0011	0.6575
p(context)	0.4265	0.4842	0.0404	0.0437	0.0052	

Joint Probabilities with marginals

	computer	**data**	**result**	**pie**	**sugar**
cherry	0	0	0	4.38	3.30
strawberry	0	0	0	4.10	5.51
digital	0.18	0.01	0	0	0
information	0.02	0.09	0.28	0	0

The PPMI Matrix