Artificial Intelligence

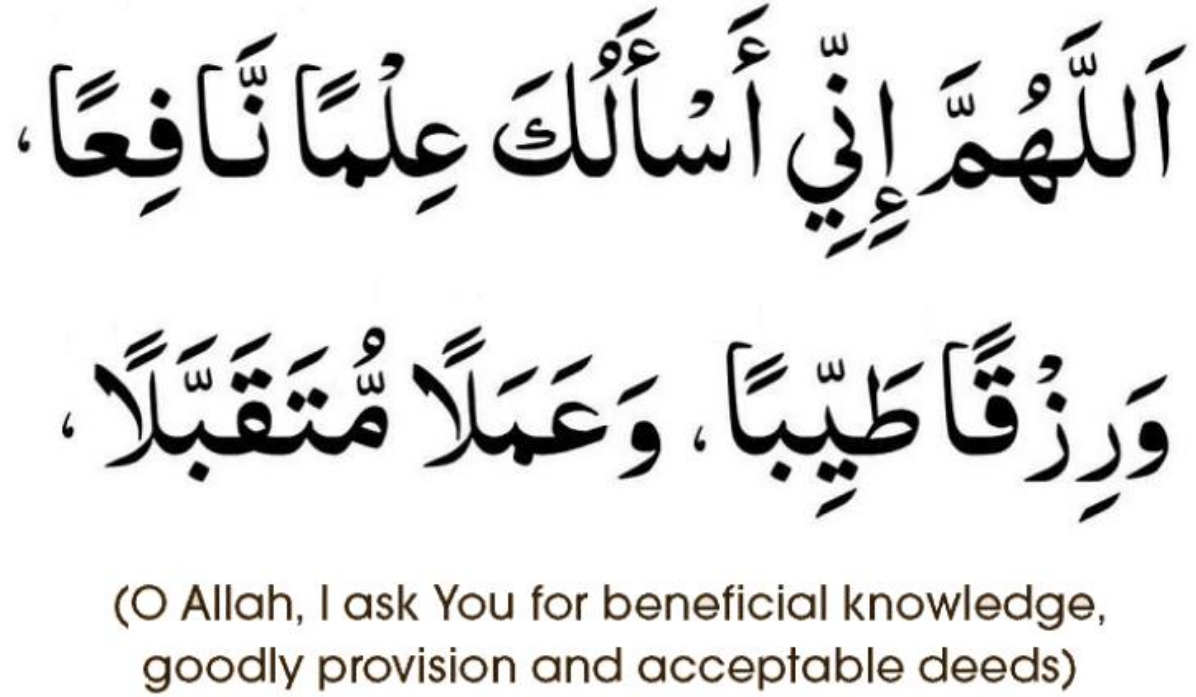
Unit 08

Word Embeddings

By:

Syeda Saleha Raza

AL NAFI,
A company with a focus on education,
wellbeing and renewable energy.



اے اللہ ، میں آپ سے سوال کرتی ہوں نفع بخش علم کا، طیّب رزق کا، اور اس عمل کا

(Sunan Ibn Majah: 925)

Acknowledgement

- A Guide to Word Embedding. What are they? How are they more useful… | by Shraddha Anala | Towards Data Science
- Generative AI exists because of the transformer

Outline

- What are word embeddings?
- How are they learnt?
- How do they help?

Turning words into numeric vectors

Bag of words

1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet

```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

Scoring

1. The sun is shining
2. The weather is sweet
3. The sun is shining and the weather is sweet

```
{'the': 5, 'shining': 2, 'weather': 6, 'sun': 3, 'is': 1, 'sweet': 4,
'and': 0}
```

```
[[0 1 1 1 0 1 0]
 [0 1 0 0 1 1 1]
 [1 2 1 1 1 2 1]]
```

Scoring: Binary, Count, TF-IDF

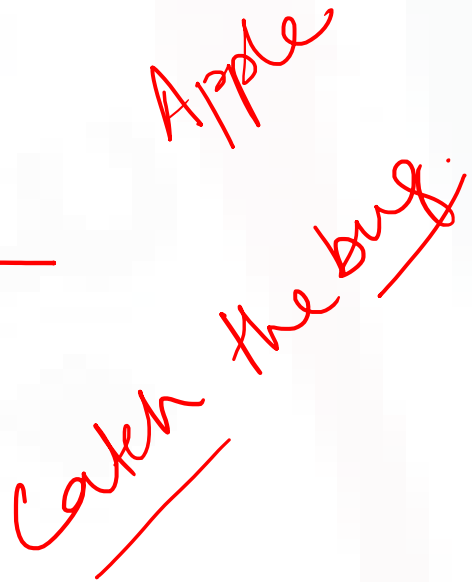
TF-IDF Score

```
[[ 0.     0.43   0.56   0.56   0.     0.43   0.   ]
 [ 0.     0.43   0.     0.     0.56   0.43   0.56]
 [ 0.4    0.48   0.31   0.31   0.31   0.48   0.31]]
```

The scores have the effect of highlighting words that are distinct (contain useful information) in a given document. Thus the idf of a rare term is high, whereas the idf of a frequent term is likely to be low.

What is missing in these vectors?

Context + Semantics



Word Embeddings

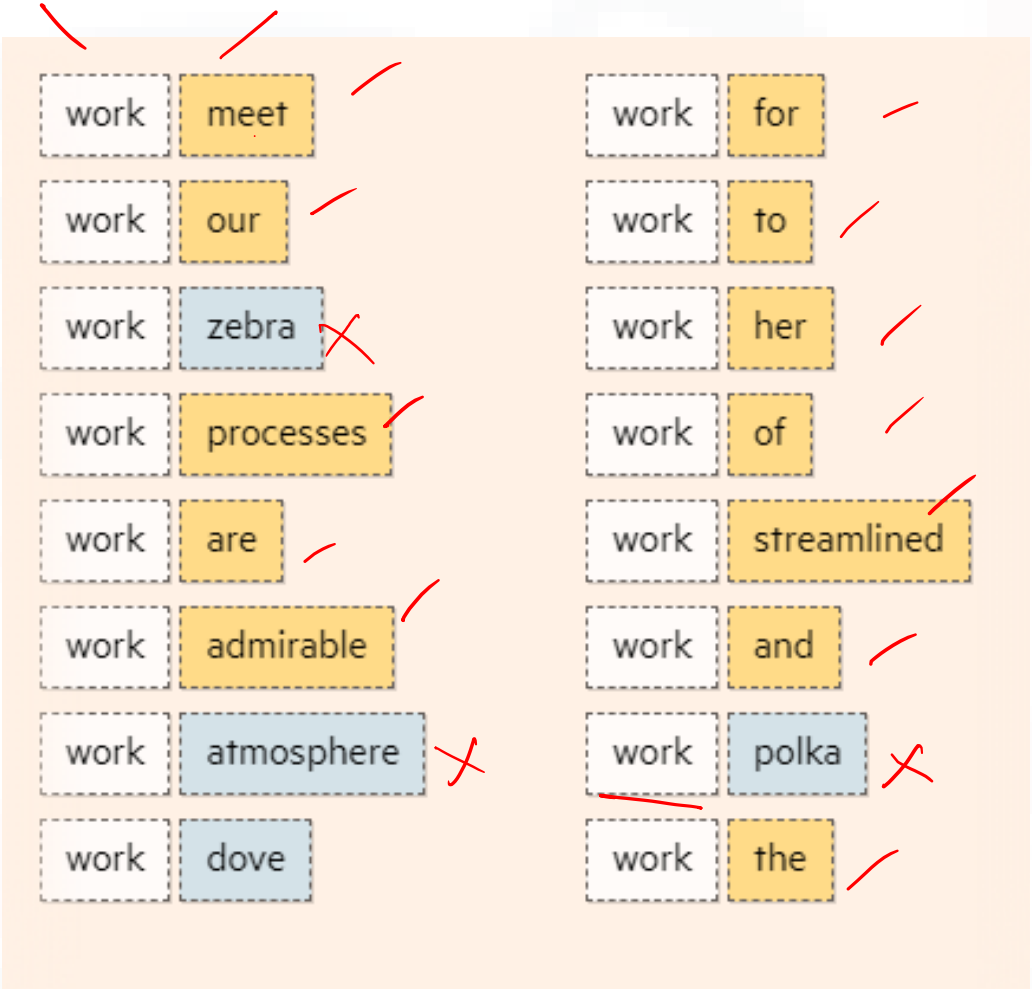
Basic Principle

"A word is known by the company it keeps."

Looking at nearby words

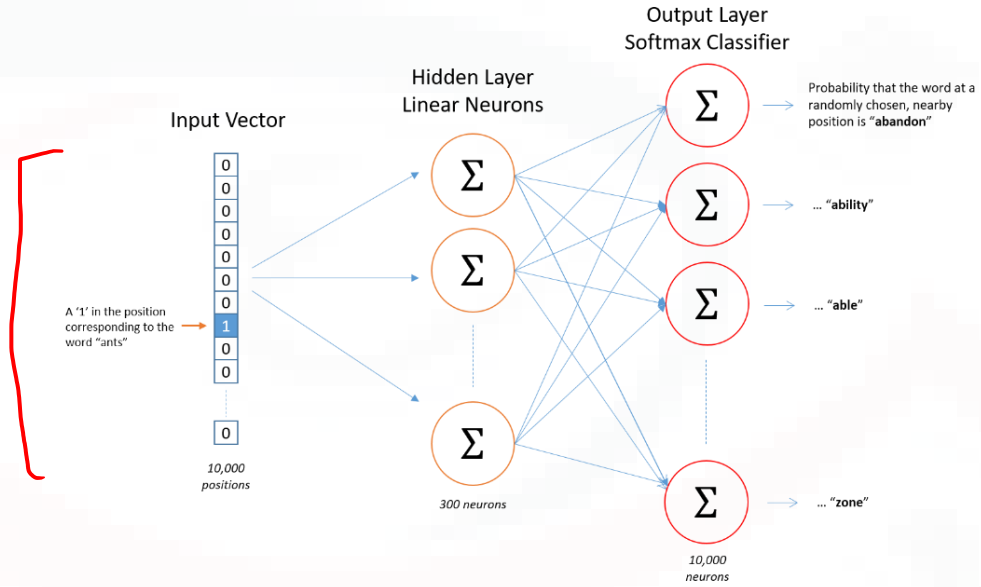
They have executed their group **work** with confidence.
My team's collaborative **work** produced outstanding results.
You should finish this **work** before the afternoon.
This is so rewarding to **work** on creative projects of this type.
The team delegated the **work** responsibilities evenly.
The artist's unique style of **work** impressed the critics.
His commitment to quality **work** has earned him recognition.
You have given her inspiration to continue her **work** amidst difficulties..
They had to adapt and **work** with limited resource.
……

Looking at nearby words

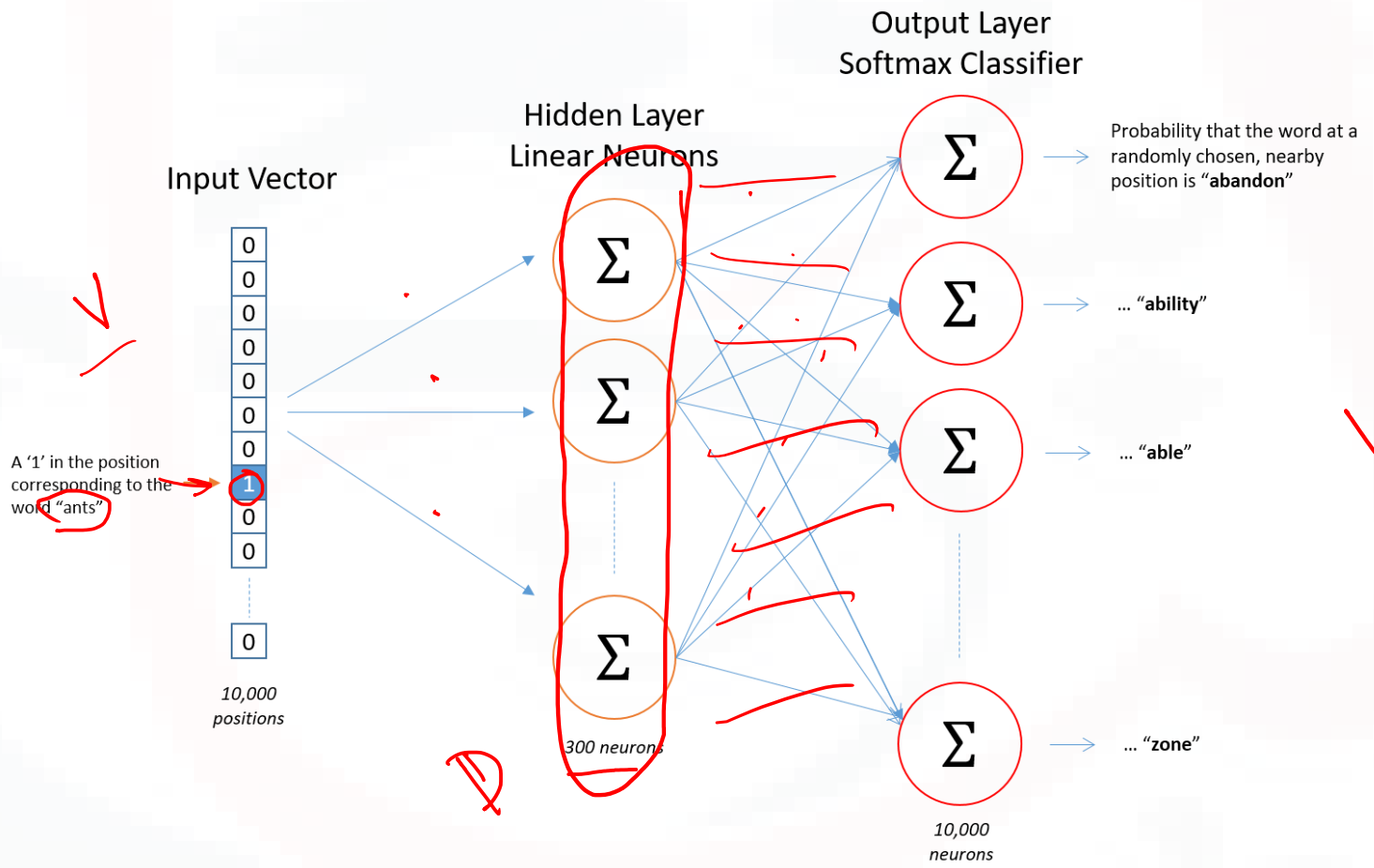


Learning Word Embeddings

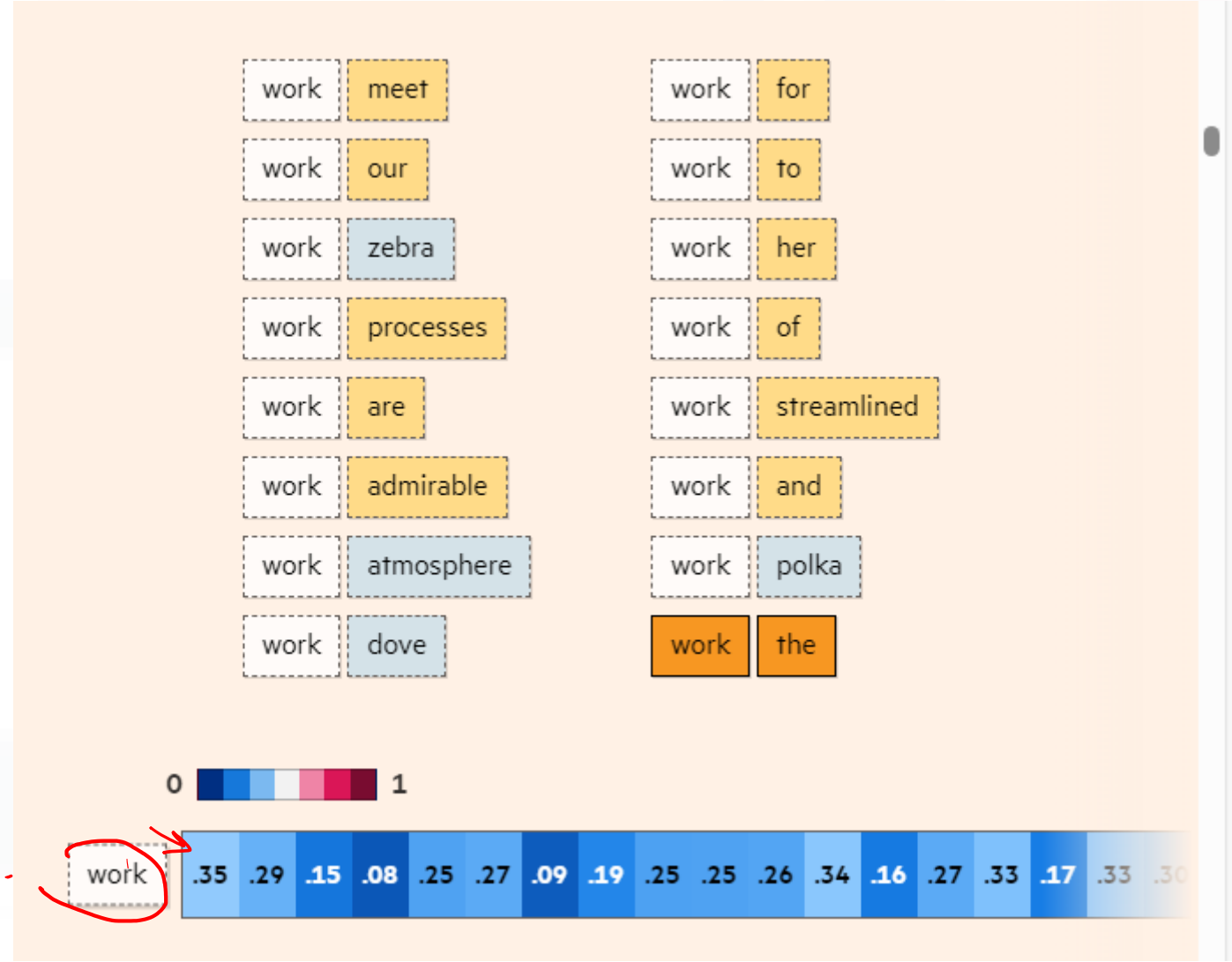
- We can train a neural network with a single hidden layer with the objective of maximizing the probability of the next words given the previous words.
- The network is not used for the task it has been trained on. The rows of the hidden layer weight matrix are used instead as the word embeddings. For a hidden layer with N=300 neurons, the weight matrix W size is V x N, where V is the size of the vocabulary set.



Learning Word Embeddings

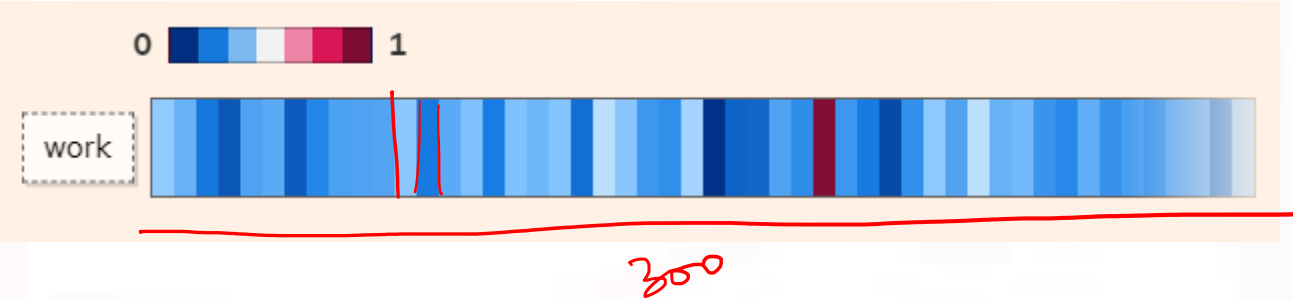


Learning word vectors



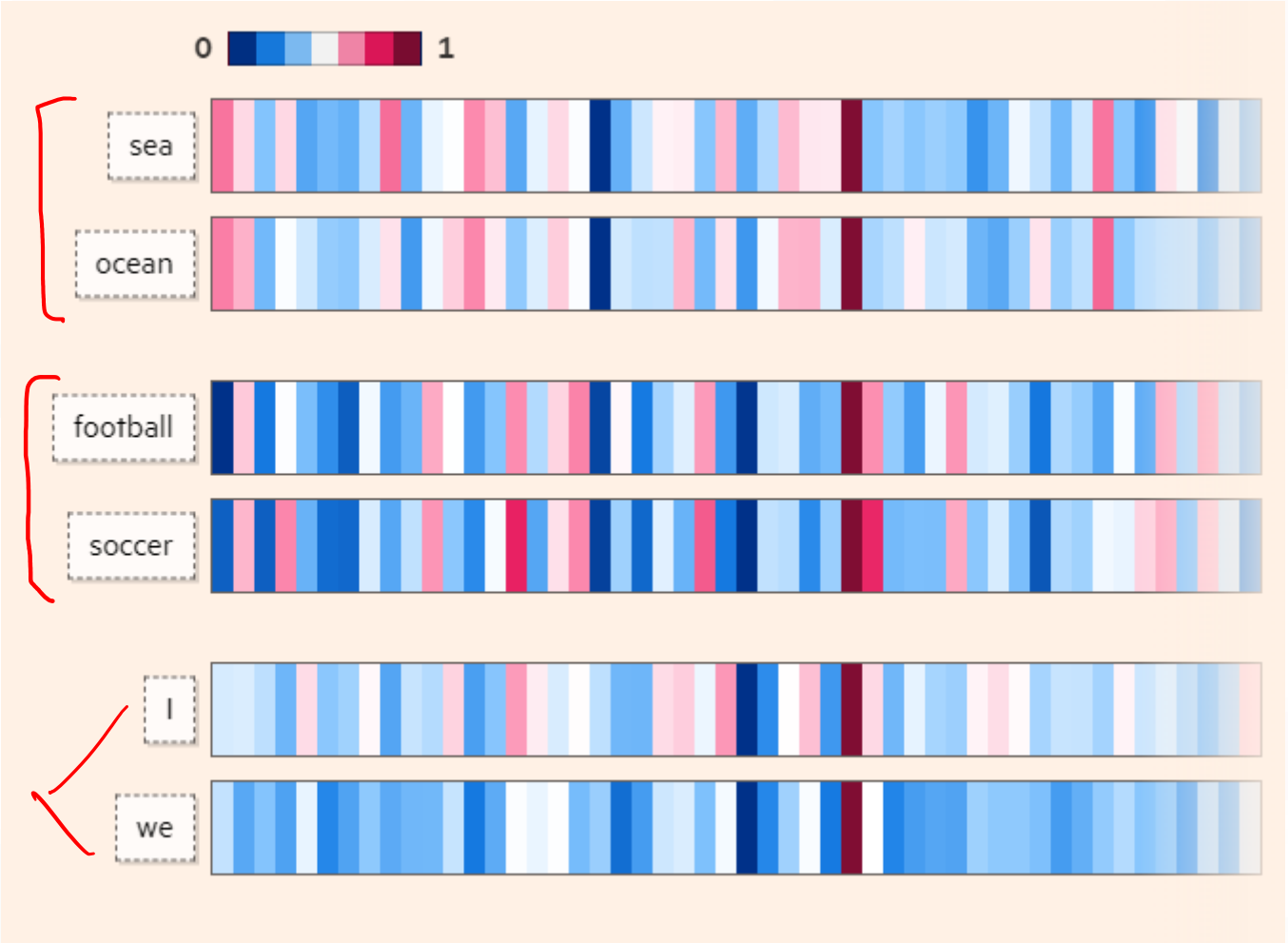
Interpreting Embeddings

Visualising Embeddings

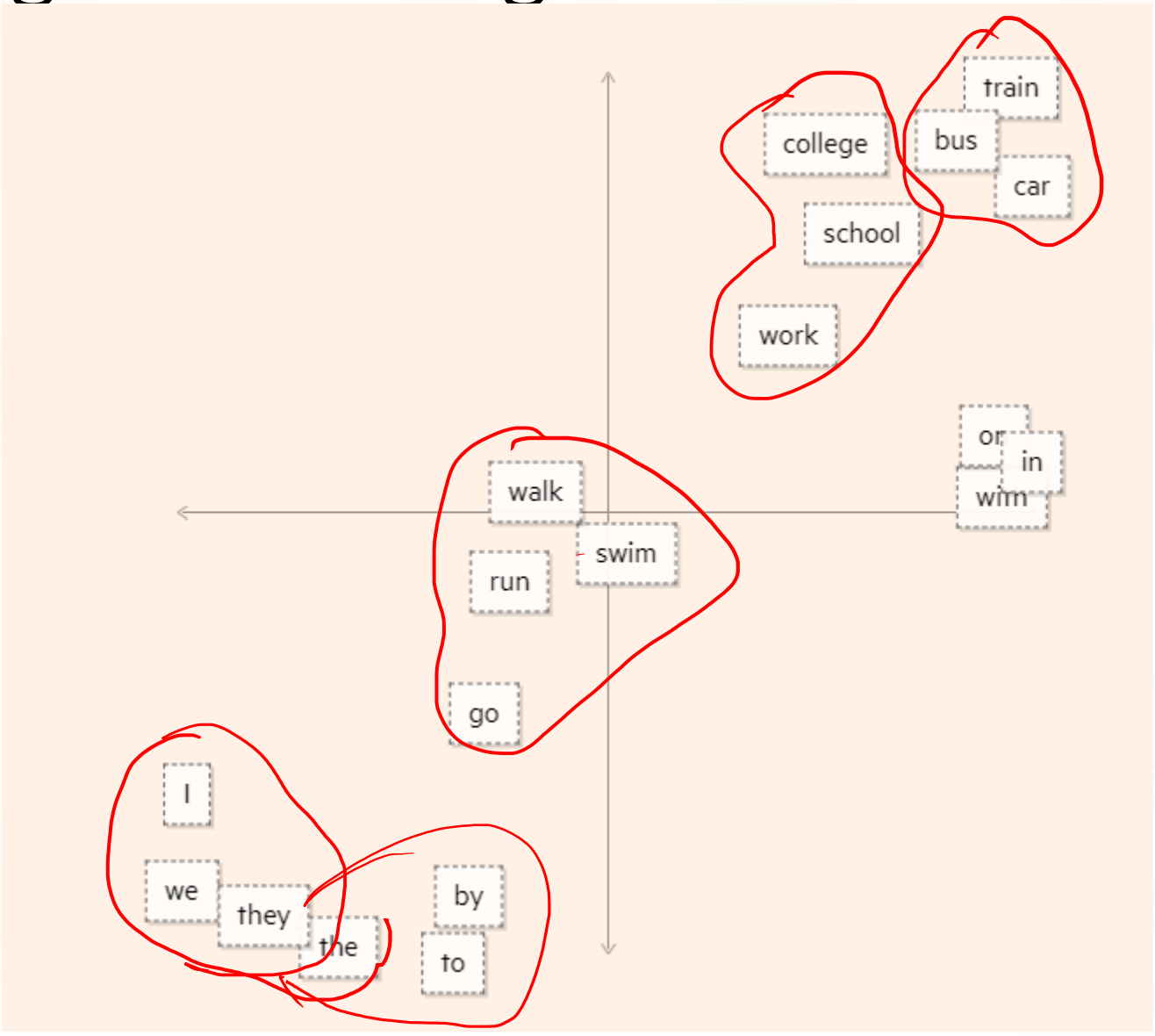


- A word embedding can have hundreds of values, each representing a different aspect of a word's meaning. Just as you might describe a house by its characteristics — type, location, bedrooms, bathrooms, storeys — the values in an embedding quantify a word's linguistic features.

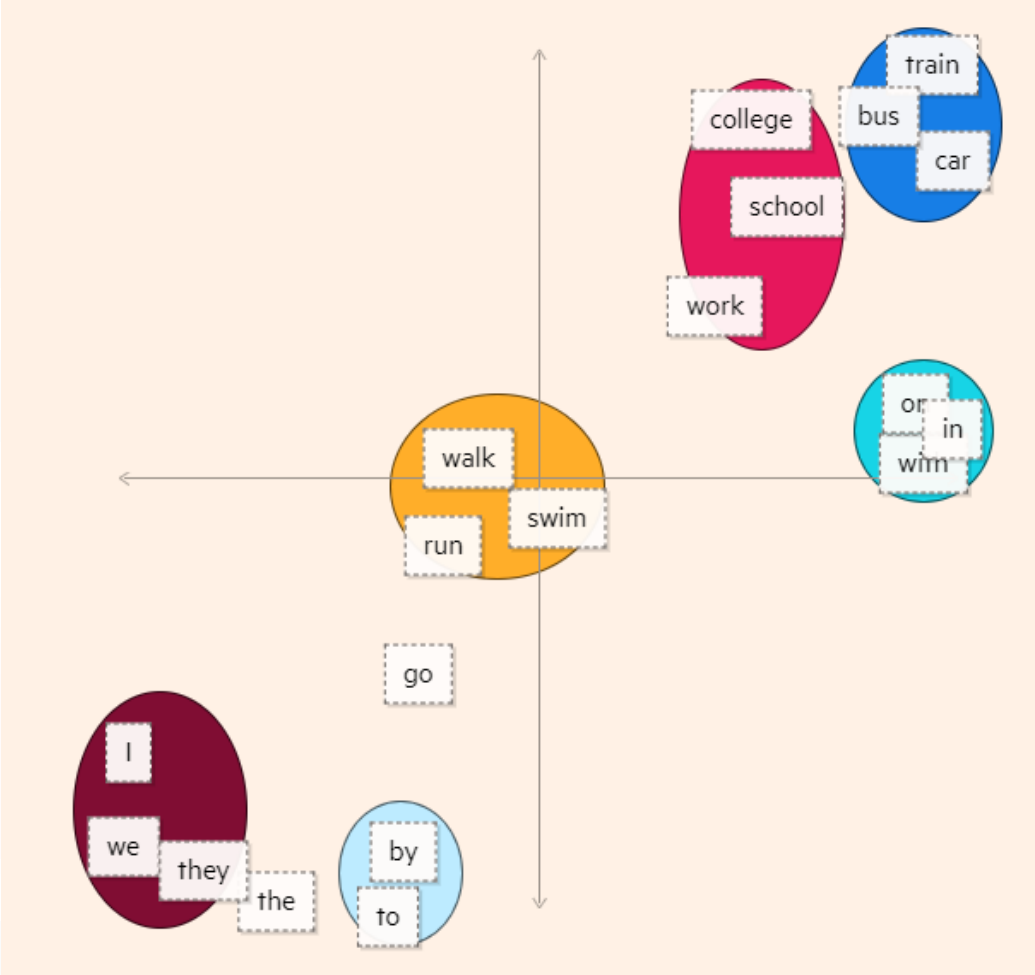
Visualising Embeddings



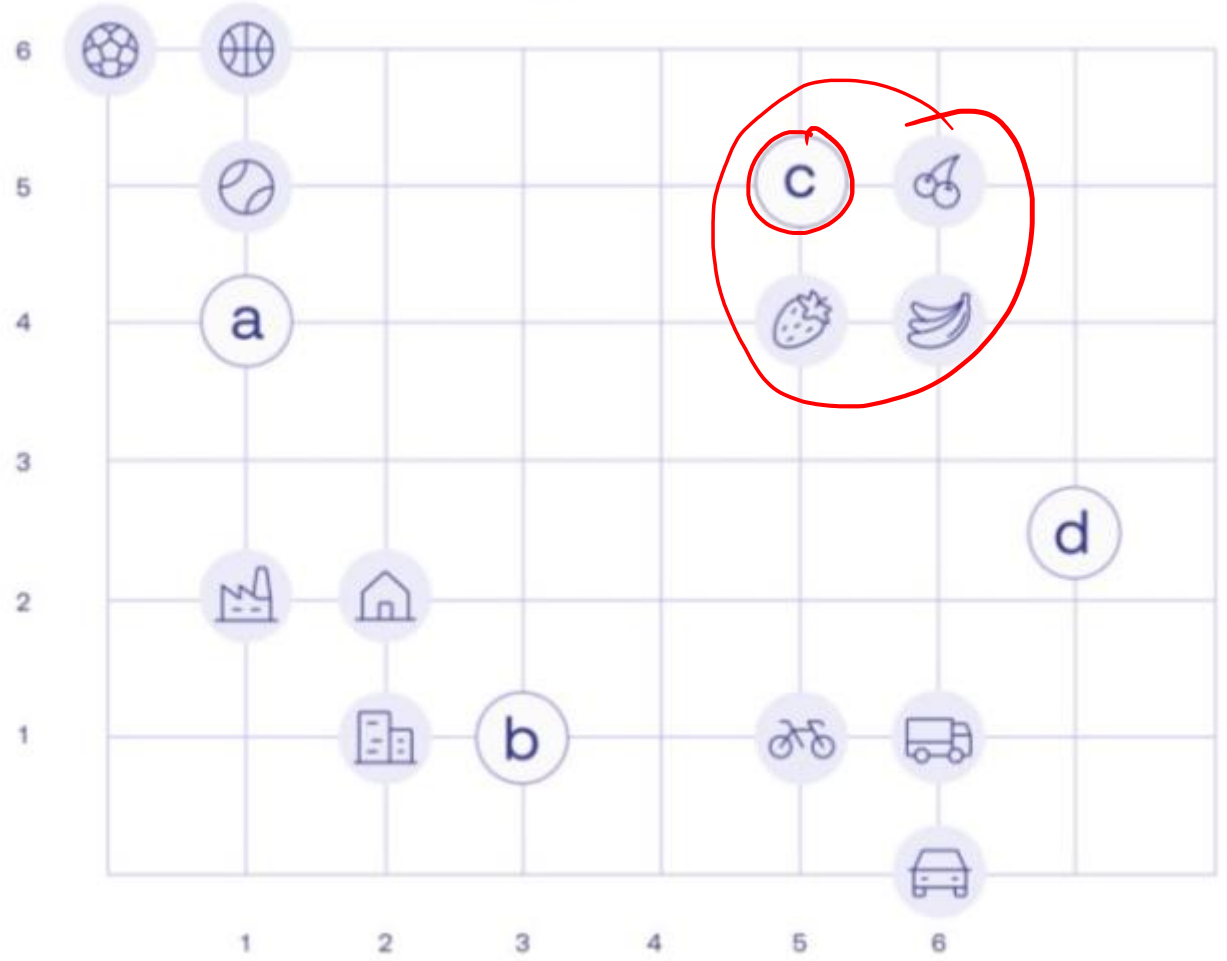
Visualising Embeddings



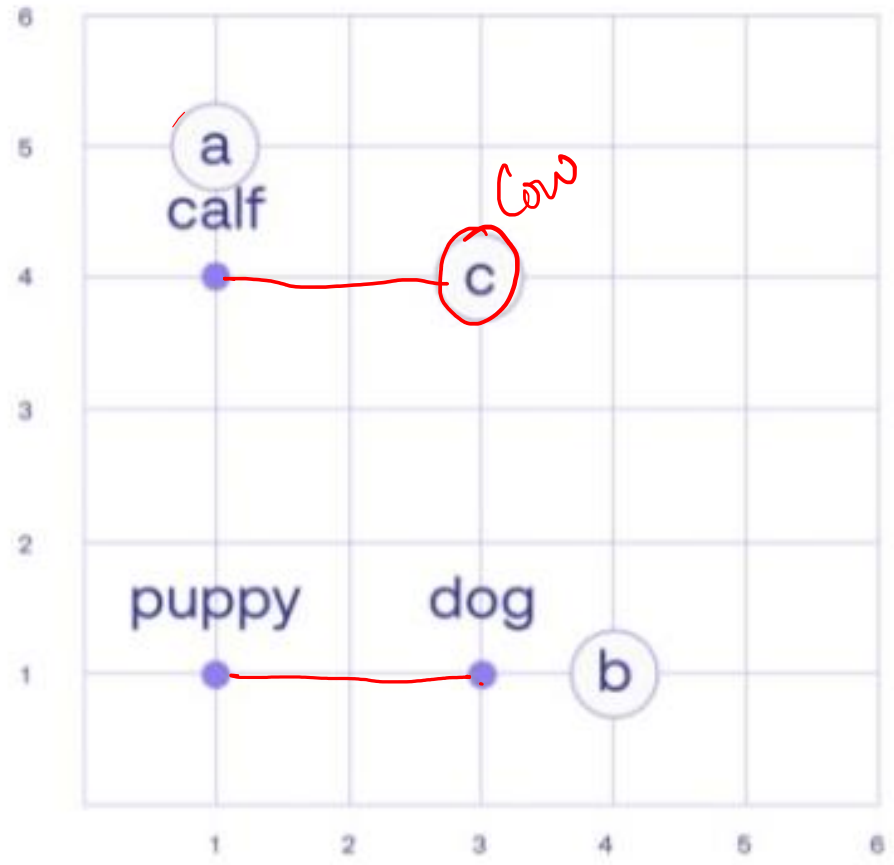
Visualising Embeddings



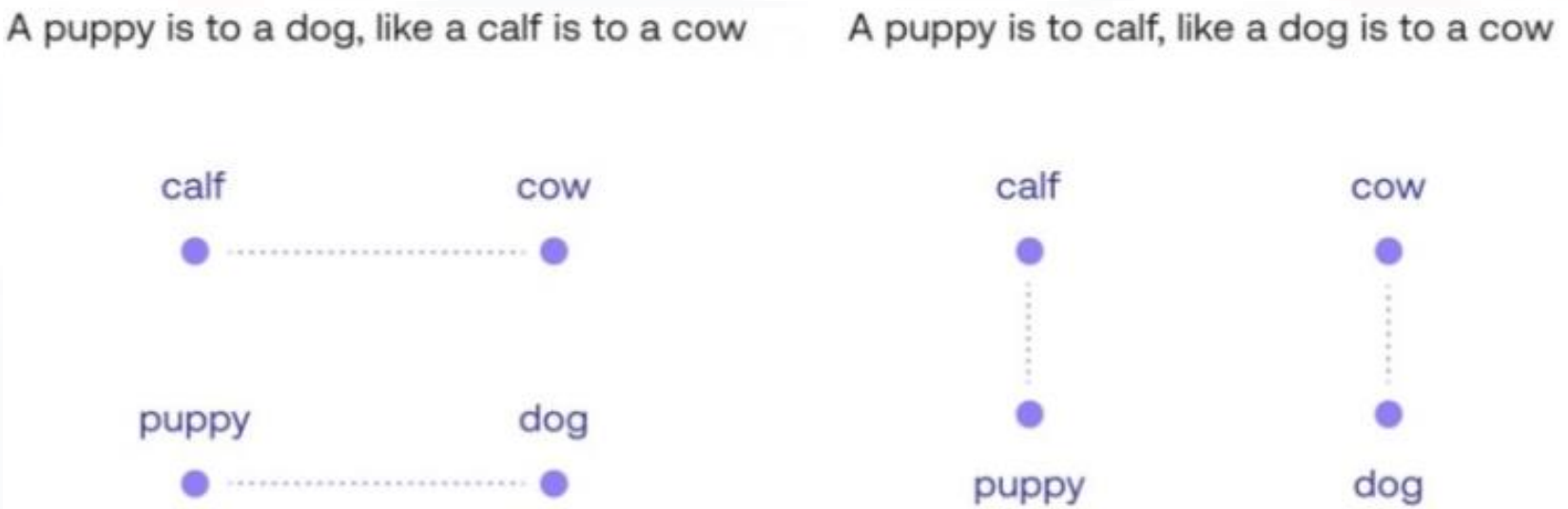
Where would you put the word ‘apple’?



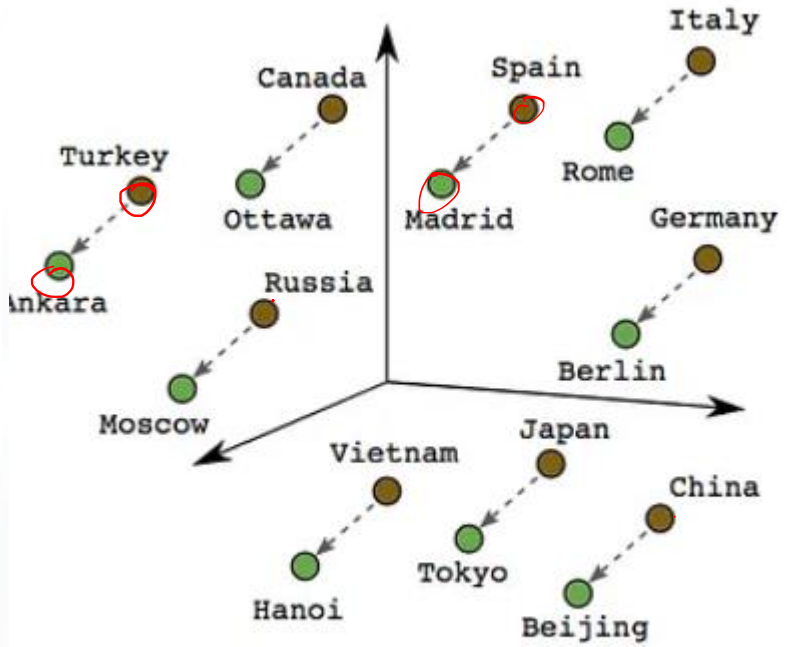
Where would you put the word ‘cow’?



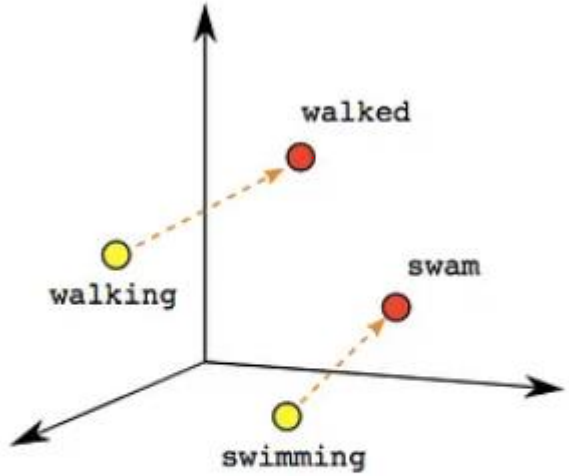
Learning relationships



Learning relationships

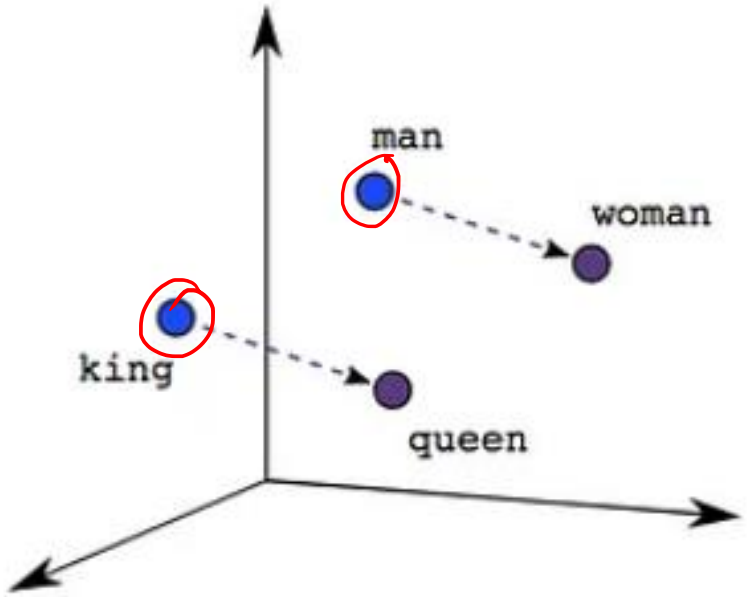


Country-Capital



Verb Tense

Learning relationships



Male-Female

Capturing Relationships

- The relationship of Man:King is the same as Woman:Queen.

vector (Man) – vector (King) + vector (Queen) = vector (Woman)

- If we build vectors for France, Italy, and Paris using Word2vec, and consider the following equation:

vector (France) + vector (Rome) - vector (Italy) = ? vector(Paris)

- The output would be vector (Paris).

Analogies in word vectors

king:queen::man:[woman, Attempted abduction, teenager, girl]
China:Taiwan::Russia:[Ukraine, Moscow, Moldova, Armenia]
house:roof::castle:[dome, bell_tower, spire, crenellations, turrets]
knee:leg::elbow:[forearm, arm, ulna_bone]
Donald Trump:Republican::Barack Obama:[Democratic, GOP, Democrats, McCain]
building:architect::software:[programmer, SecurityCenter, WinPcap]

https://medium.com/datadriveninvestor/solving-analogies-using-word2vec-13b9e01eca1

Pre-trained embeddings

Word2Vec

- The pre-trained Google Word2Vec model was trained on Google news data which has around 100 billion words. It contains 3 million words and phrases and was fit using 300-dimensional word vectors.
- The embeddings can be downloaded from :

  GoogleNews-vectors-negative300.bin.gz - Google Drive

gensim

Glove

- Stanford researchers also have their own word embedding algorithm called Global Vectors for Word Representation (GloVe).
- The training corporal had 6 Billion tokens and the final embeddings have 400K words. The vectors are available in 50, 100, 200 and 300 dimensions. The vectors are trained on Wikipedia data and can be downloaded from:

  http://nlp.stanford.edu/data/glove.6B.zip OR

  GloVe: Global Vectors for Word Representation (stanford.edu)

Code Demo

- Using pre-trained word embeddings (word2vec, glove)
  - Learning word embeddings on our own text corpus

جزاك اللهُ

To ask questions, Please use communities link (for respective course) within portal

https://portal.alnafi.com/enrollments