

Natural Language Processing

Word Embeddings and Semantic Similarity

Words as numbers


Computers can only work with numbers. Individual characters are stored as numbers ('A' is character code 65 for instance)
For computers to process words we need to find a way to store them as numbers too

Options

1. Store each word as a list of its character codes: **Ant** would be [65, 78, 84]
2. Give each word in the whole vocabulary its own number.

Vocabulary

index:	word:
0	aardvark
1	able
...	...
5281	is
5282	island
...	...
9999	zombie



**10,000
words
with
indices**

These both work, but they don't tell the computer anything about the word. Can we do better?

How about a system where words with similar meaning have numbers that are somehow 'closer' to each other?

Encoding words as dimensions

Take 5 example words from our vocabulary:

“aardvark”, “black”, “cat”, “duvet” and “zombie”

As humans we know that words all are objects with meaning.

Let’s hand-craft some semantic features for these 5 words.

Specifically, let’s represent each word as having some sort of value between 0 and 1 for four semantic qualities, “animal”, “fluffiness”, “dangerous”, and “spooky”:

	animal	fluffiness	dangerous	spooky
aardvark	0.97	0.03	0.15	0.04
black	0.07	0.01	0.20	0.95
cat	0.98	0.98	0.45	0.35
duvet	0.01	0.84	0.12	0.02
zombie	0.74	0.05	0.98	0.93

Now we can think about these features as dimensions and can plot them (only 2d or 3d though - 4d is hard!!)

Notice that now we can talk about how close words are in a specific dimension:

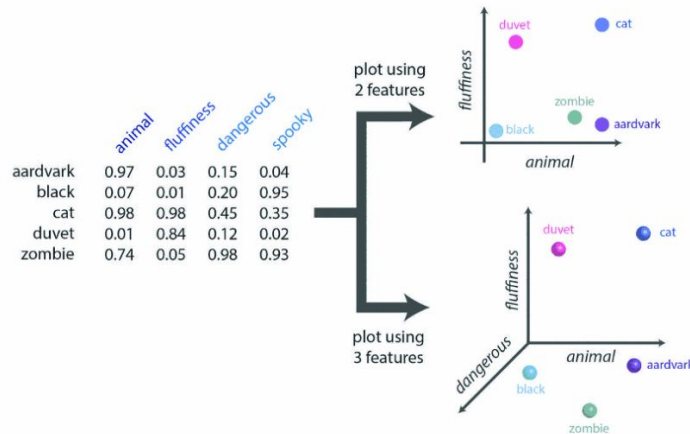
cat and aardvark are both very ‘animality’

duvet and cat aboth very ‘fluffy’

black and zombie are both very ‘spooky’

In fact there are mathematical operations (e.g. cosine similarity) that let us compare similarity across **all** of the dimensions at once

So... if we can describe all the features of all words then we can use maths to figure out how similar or far apart they are from each other!!!



Machine learning for word embeddings

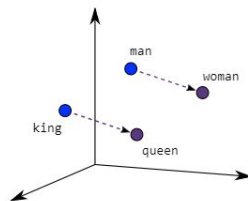
“Word embeddings” are just a fancy name for a list of numbers that describe the word. On the previous slide our embeddings for **zombie** were: $[0.74, 0.05, 0.98, 0.93]$

Machine learning researchers have trained neural networks on huge amounts of text (billions of words) to ‘learn’ embeddings. For each word in a sentence they use the surrounding words to learn what it means.

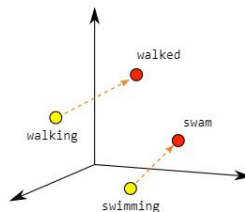
The result is a set of embeddings that we can all use (for free) to represent and do maths with words

It’s not only similarity measures, you can even add and subtract words from each other!!

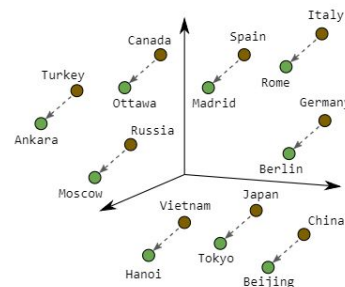
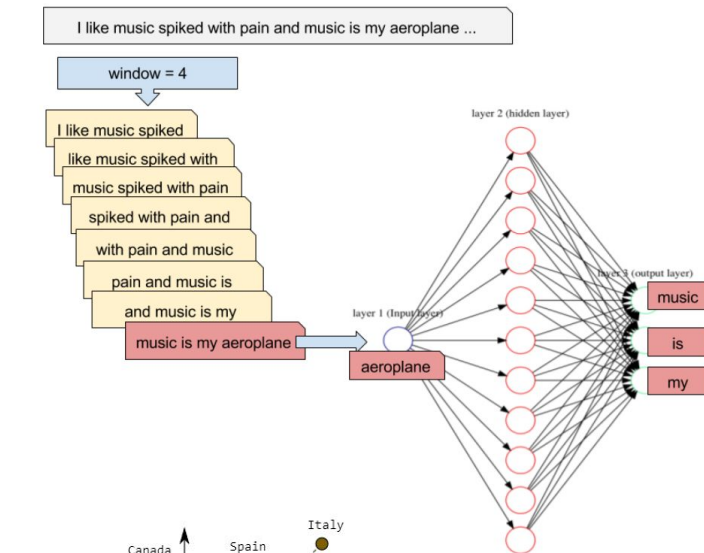
King - Man + Woman = Queen



Male-Female



Verb Tense



Country-Capital