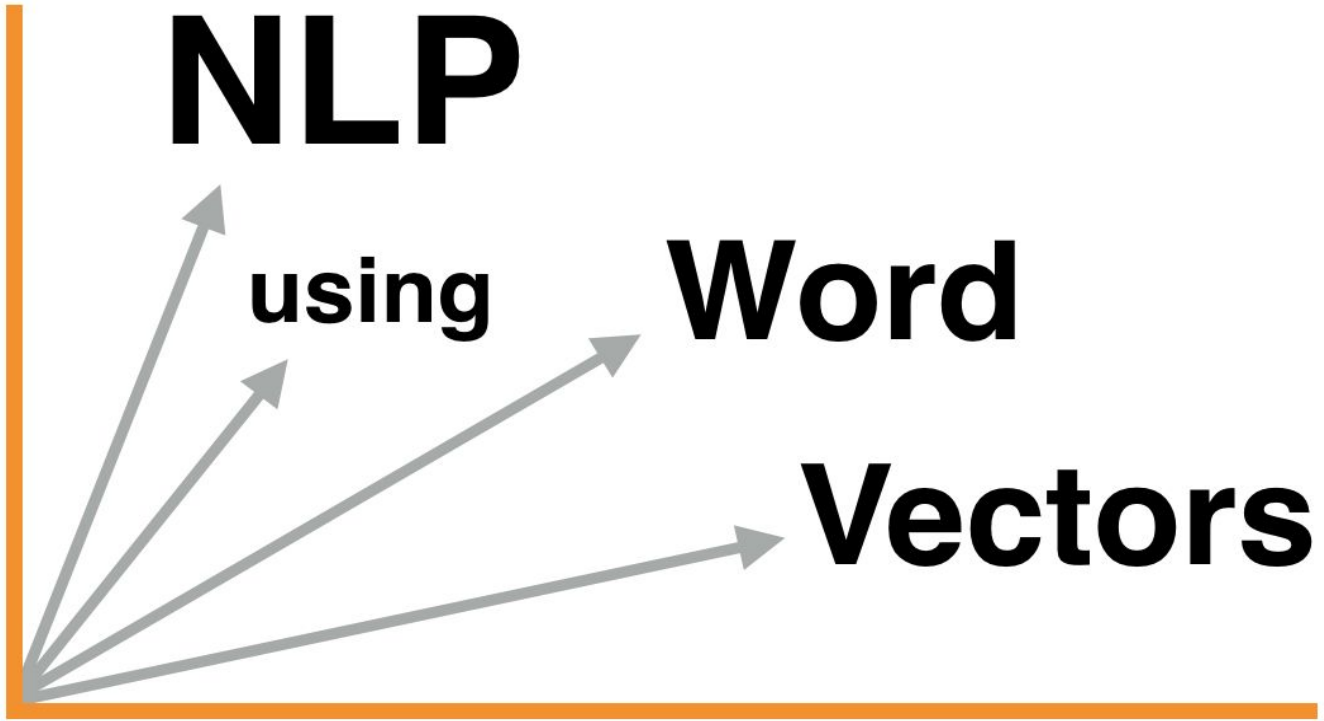


#CLDSPN

@ZackAkil

Why do text encoding?

“CAT” → 36

“CAR” → 179

“BOAT” → 40

$y = 5 * X$

$y = 5 * \text{“CAR”}$ 🤔

$y = 5 * 179$ ✅

@ZackAkil

One-hot-encoding (and word-count-vectors)

1. 'the cat sat on the car'

2. 'the data is on the boat'

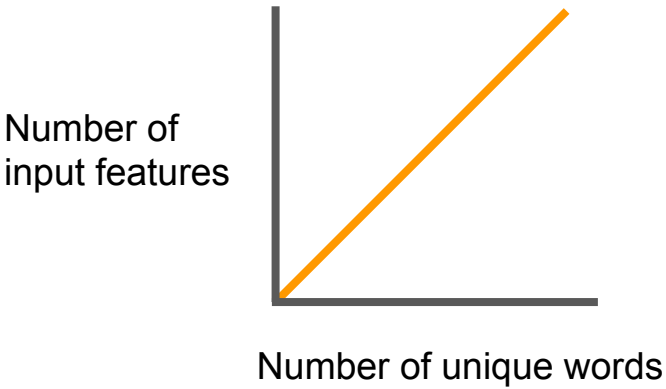
3. 'why is my cat on the boat'

#	CAT	CAR	BOAT	SAT	DATA	ON	THE	IS	WHY	MY
1	1	1	0	1	0	1	2	0	0	0

@ZackAkil

Issue with one-hot-encoding (and word-count-vectors)

Dimensionality



Word Distance

Document	CAT	CAR	BOAT
‘Cat’	1	0	0
‘Car’	0	1	0
‘Boat’	0	0	1

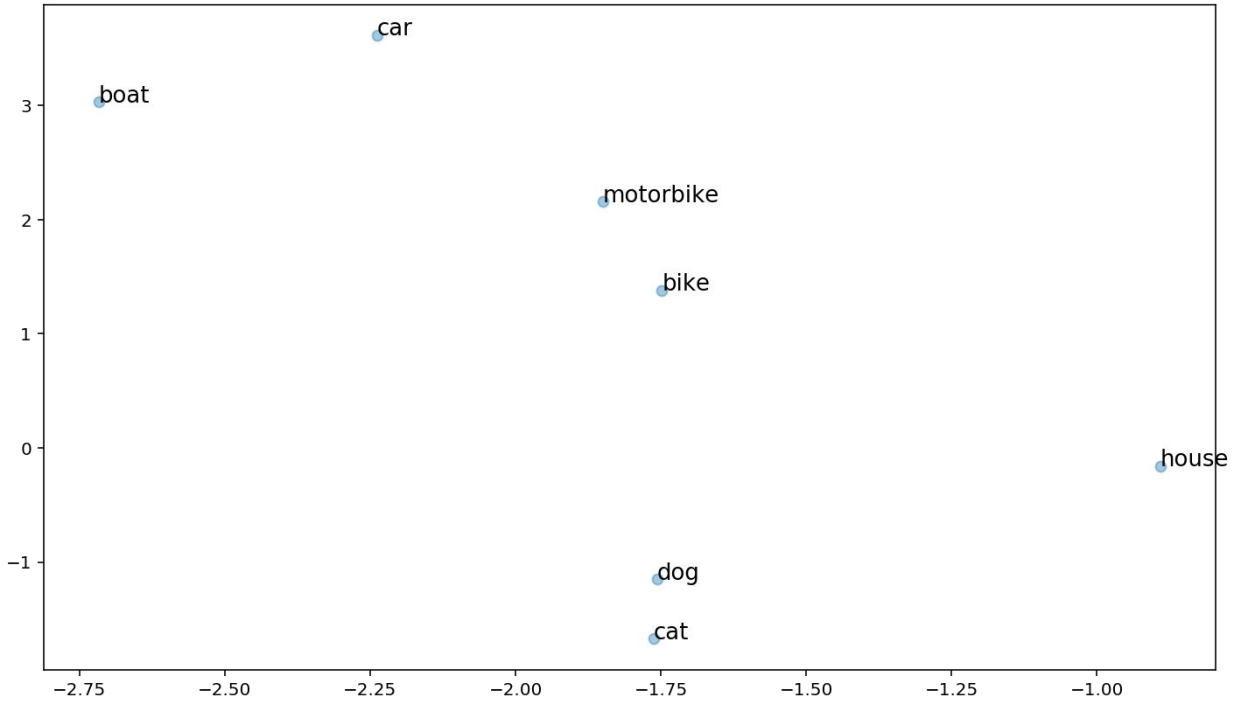
@ZackAkil

Word vector embeddings

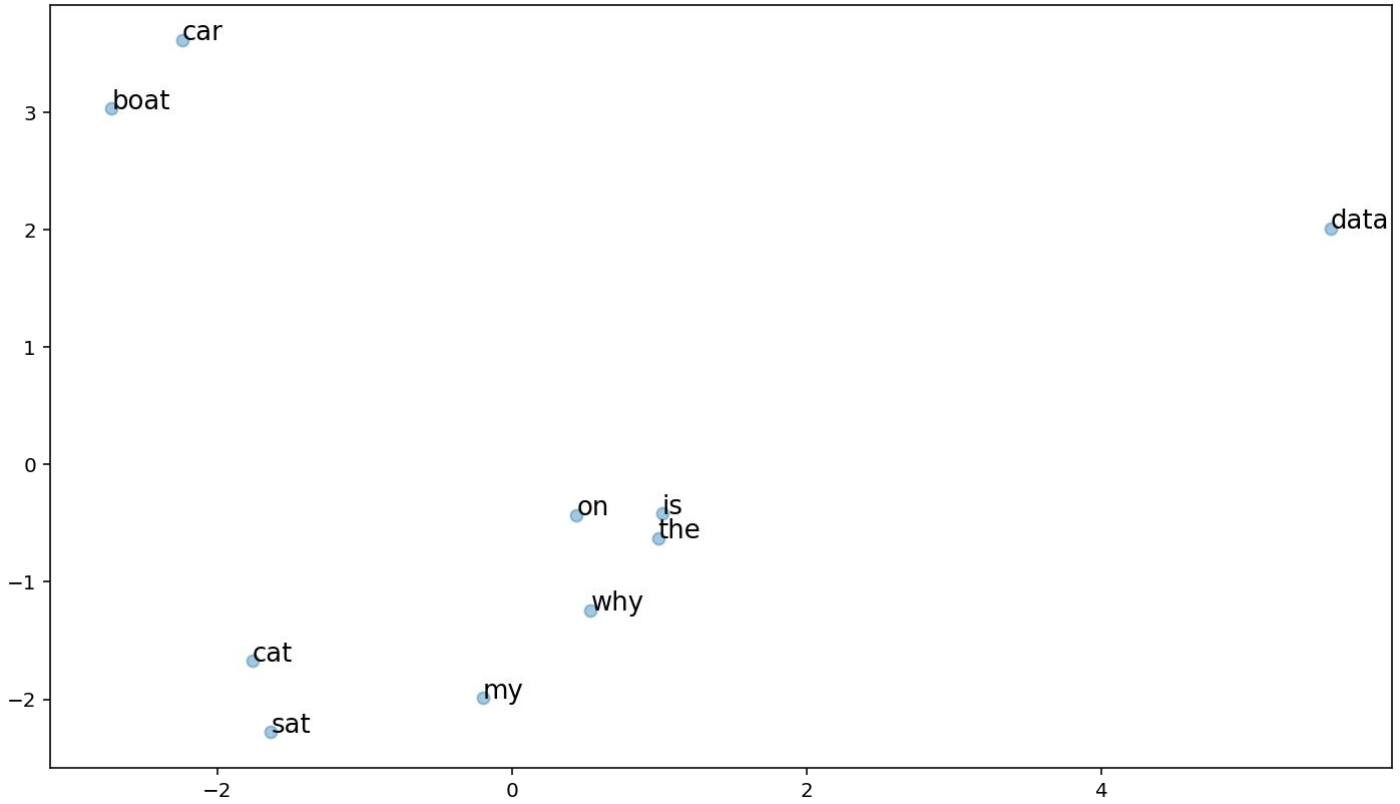
Document	Size	Cuteness	Transport	Wheels	Cargo
‘Cat’	0.1	0.9	0.3	0.1	0.0

@ZackAkil

Word vector power

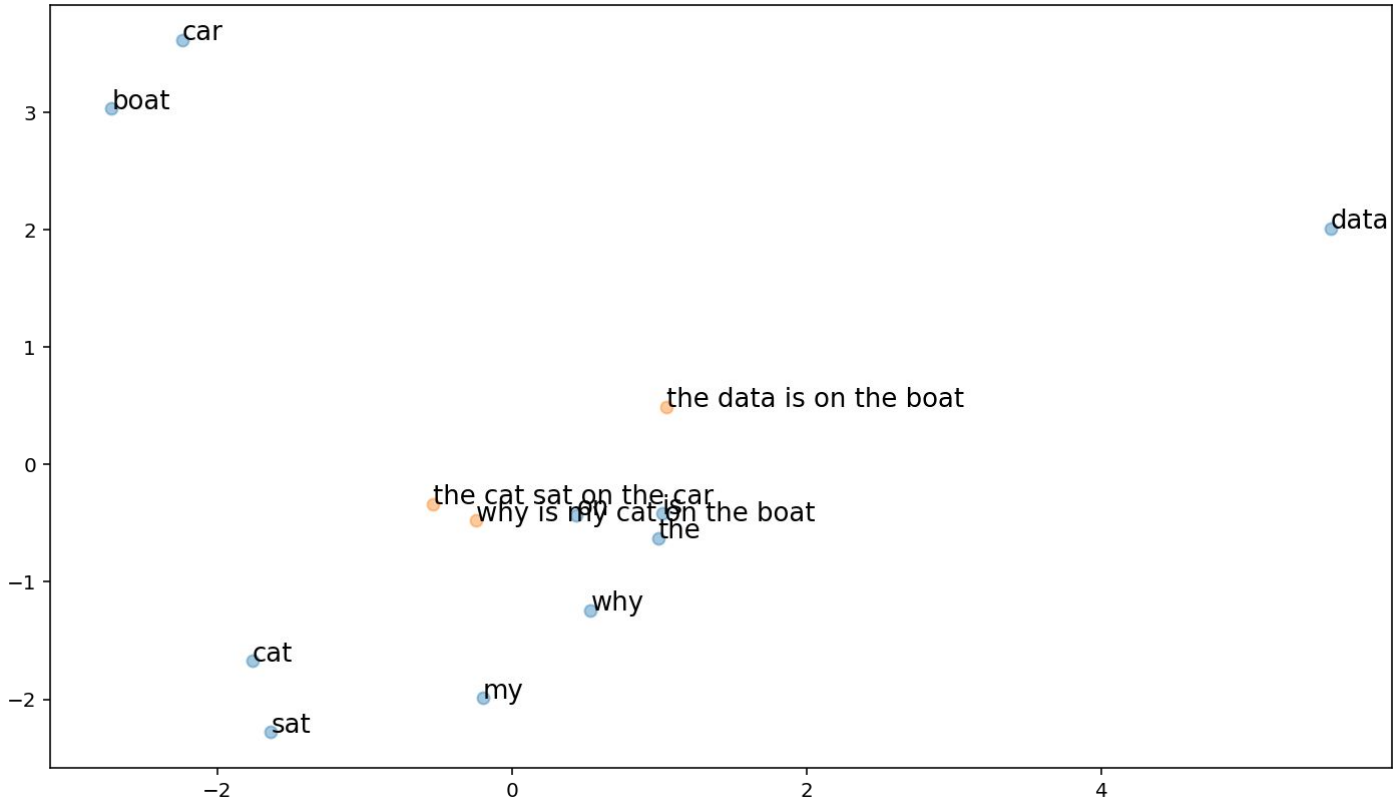


@ZackAkil



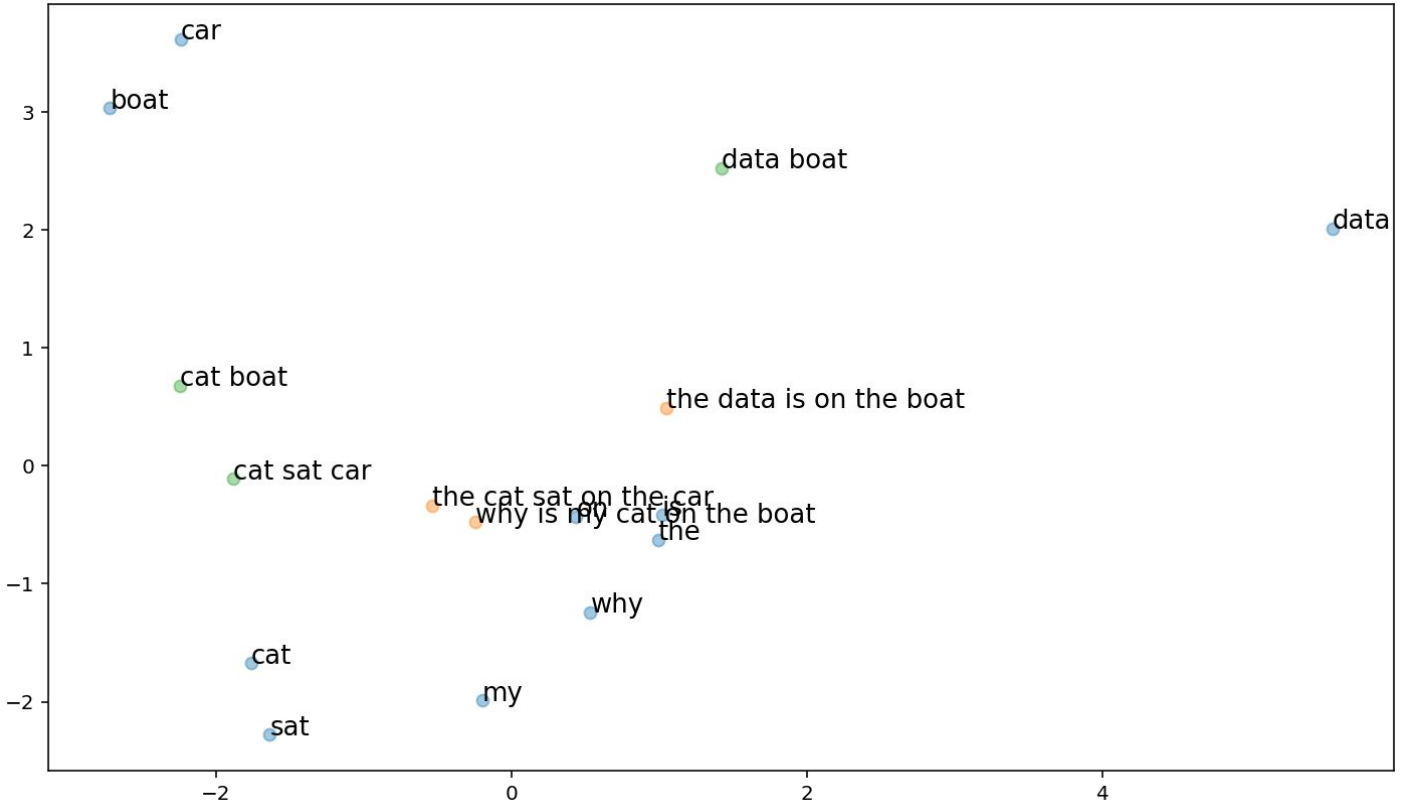
@ZackAkil

Document vector



@ZackAkil

Stop word removal



@ZackAkil