

Word representation in machine learning problem

1. Localist representation: One-hot encoding vector, a vector of zeros, excepts the position where the target is in the word pool.
2. **Words vector**: word embeddings or word representations, which are distributed representation. Captures word meanings by a vector of real valued numbers (as opposed to dummy numbers) where each point captures a dimension of the word's meaning and where semantically similar words have similar vectors.

		Dimensions					
Word vectors	dog	-0.4	0.37	0.02	-0.34	animal	
	cat	-0.15	-0.02	-0.23	-0.23	domesticated	
	lion	0.19	-0.4	0.35	-0.48	pet	
	tiger	-0.08	0.31	0.56	0.07	fluffy	
	elephant	-0.04	-0.09	0.11	-0.06		
	cheetah	0.27	-0.28	-0.2	-0.43		
	monkey	-0.02	-0.67	-0.21	-0.48		
	rabbit	-0.04	-0.3	-0.18	-0.47		
	mouse	0.09	-0.46	-0.35	-0.24		
	rat	0.21	-0.48	-0.56	-0.37		

Advantage:

- Relatively smaller dimension
- similar words as similar word vectors, and can be measured mathematically.
- support mathematical operation e.g. King - Man + Women = Queen

How to construct word representation?

Word2vec

\$ \int_a \$