Modelo espaço vetorial

- Cada sentença/documento é representada em um vetor, onde cada posição do vetor corresponde a uma palavra (termo) no documento.

- Nessa representação, as palavras são consideradas independentes, formando um conjunto desordenado em que a ordem de ocorrência das palavras não importa.

Modelo espaço vetorial

	t_1	t_2	...	t_M
d_1	a_{11}	a_{12}	...	a_{1M}
d_2	a_{21}	a_{22}	...	a_{2M}
⋮	⋮	⋮	⋱	⋮
d_N	a_{N1}	a_{N2}	...	a_{NM}

- Geralmente *anm* é obtido de 2 formas:
  - (1) um valor que indica se um termo está ou não no documento
  - (2) um valor que indica a importância (distribuição) do termo ao longo da coleção.

- Representação **esparsa** > matriz de **alta dimensionalidade**
  - No vetor de cada documento são representadas todas as palavras da coleção, e não somente aquelas presentes no documento.

Problema

- Geralmente, o desempenho dos algoritmos é prejudicado com dados esparsos e pela alta dimensionalidade
- A grande dimensionalidade provoca um alto custo computacional, tornando a execução dos algoritmos muito lenta e até inviável
- Além disso, com BOW (bag of words) não sabemos nada sobre a semântica das palavras
- Mesmo com uso de stemmer para reduzir a dimensionalidade, há palavras altamente relacionados que não possuem o mesmo stem

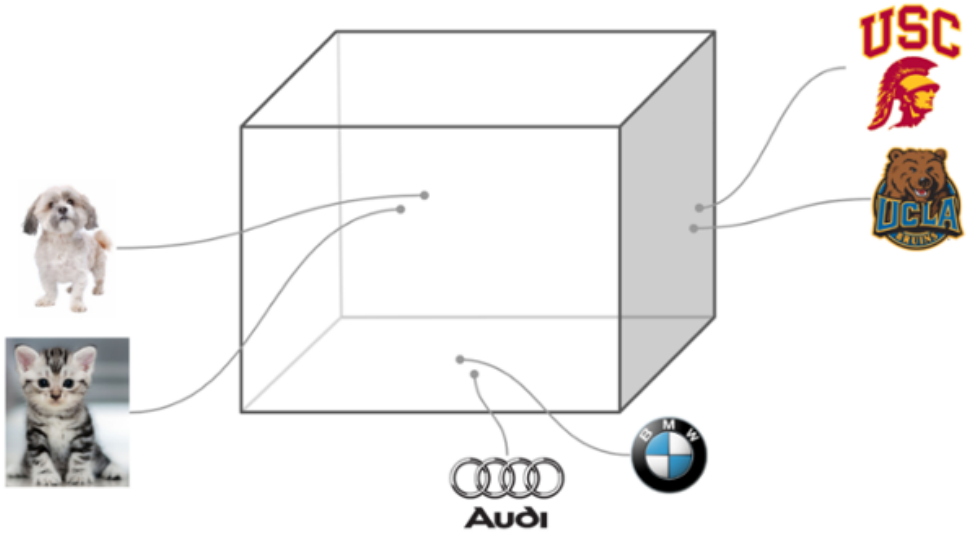
Word embeddings

- Uma maneira de representar documentos textuais numericamente é mapeando palavras em vetores de valores reais com **diversas dimensões**
  - Ou seja, uma representação distribuída de palavras na qual os vetores preservam o significado semântico e sintático das palavras em uma sentença

- Os vetores resultantes dessa representação são chamados **word embeddings** (word vectors)

- São vetores **densos** (redução da dimensionalidade) e representam **similaridade contextual** (representação mais expressiva).

Word embeddings



- *Cat* and *dog:* animais fofos, podem ser animais de estimação, tem 2 olhos, 4 pernas e 1 nariz
- *Audi* and *BMW:* companhias alemãs de automóveis de luxo
- *USC* and *UCLA*: universidades importantes em Los Angeles

Word embeddings

- Operações matemáticas podem ser feitas a partir de word vectors.
  - Exemplo: king – man + woman = queen

- Os números no vetor de palavras representam o peso distribuído da palavra entre as dimensões.

- Cada dimensão representa um significado e o peso de uma palavra nessa dimensão captura a proximidade de sua associação com e para esse significado.

Word embeddings

Dimensions

Word vectors					
	dog	-0.4	0.37	0.02	-0.34
	cat	-0.15	-0.02	-0.23	-0.23
	lion	0.19	-0.4	0.35	-0.48
	tiger	-0.08	0.31	0.56	0.07
	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

- Cada dimensão captura um significado bem definido.
  - A 1ª. dimensão representa o significado ou conceito de “animal”, o peso de cada palavra nessa dimensão representa o quanto ela se relaciona com esse conceito.

Referências

- https://nlpforhackers.io/word-embeddings/
- https://www.quora.com/What-is-word-embedding-in-deep-learning?1541294994_kis_cup_C6FA3ED5_6D17_47D1_B6E2_F4B02CC905E0_
- https://medium.com/@jayeshbahire/introduction-to-word-vectors-ea1d4e4b84bf