DEEP LEARNING INSTITUTE

LABORATÓRIO DE INTELIGÊNCIA COMPUTACIONAL APLICADA

Fundamentos de Deep Learning para Processamento de Linguagem Natural Parte 1

Dr. Cristian E. Muñoz Villalobos

WORD EMBEDDINGS

SUMÁRIO

1. NLP clássica e Bag-of-Words
2. Representação Distribuída
3. Word2Vec algorithm

MODELOS CLÁSSICOS DE NLP

Como a PNL era feita antes do Deep Learning?

- HMMs, CRFs e outros tipos de modelos como PGM
- Bag of words – um feature por palavra

  the cat sat on the mat

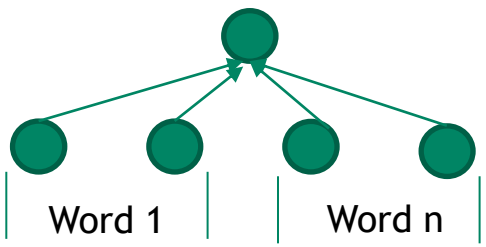
cat	sat	on	the	mat	quickly
1	1	1	2	1	0

... |Vocabulary|

- Varias formas de escolher o valor: Binário, Contagem, TF-IDF

PROBLEMAS COM REPRESENTAÇÃO BOW

Entrada esparsa (1-hot)



p >> n (overfitting!)

Sem generalização semântica

dog: 1 0 0 0 0 … 0

cat: 0 0 1 0 0 … 0

muitos dados necessários,
baixa acurácia

REPRESENTAÇÃO DISTRIBUIDA

HIPÓTESE DISTRIBUTIVA (FIRTH, 1957)

‘Você pode dizer a palavra, pela companhia que ela tem’

The cat sat on the mat

The dog sat on the mat

The elephant sat on the mat

The quickly sat on the mat

PROPIEDADES DE *EMBEDDINGS*

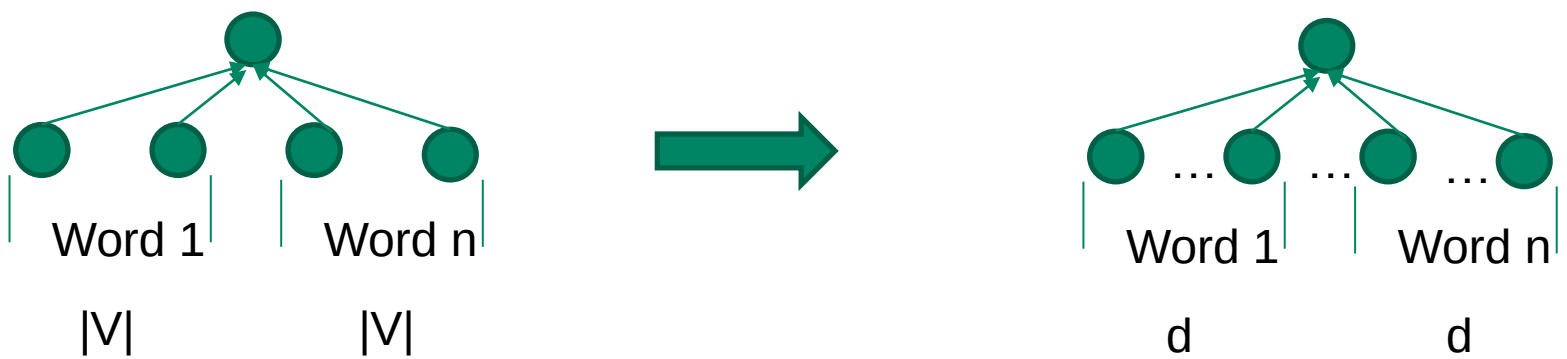
- São densos: 0.53, 0.2, -1.2, ….
- Sãi de baixa dimensão: (50 <= d <= 300)
- Incorporam a semântica do domínio:

  `king' - `man` + `woman` ≈ `queen`

  `paris` - `france` + `spain` ≈ `madrid`
- Generaliza facilmente!

PROPIEDADES DE *EMBEDDINGS* (CONT.)

São utilizados como entrada para outras tarefa de PLN:

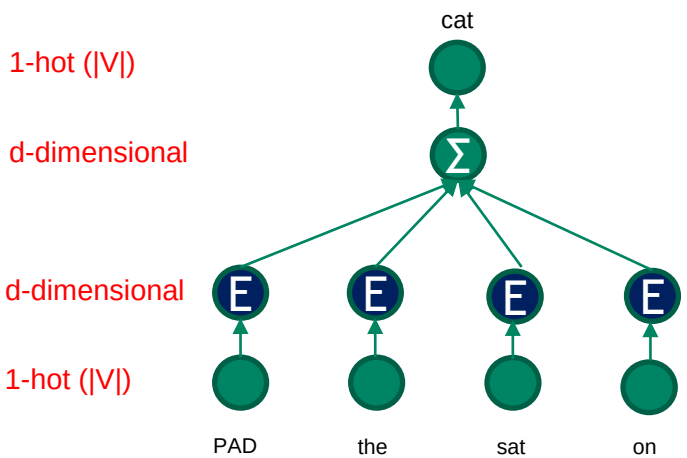


# total de parâmetros de entrada:	n * |V|	>>	n * d

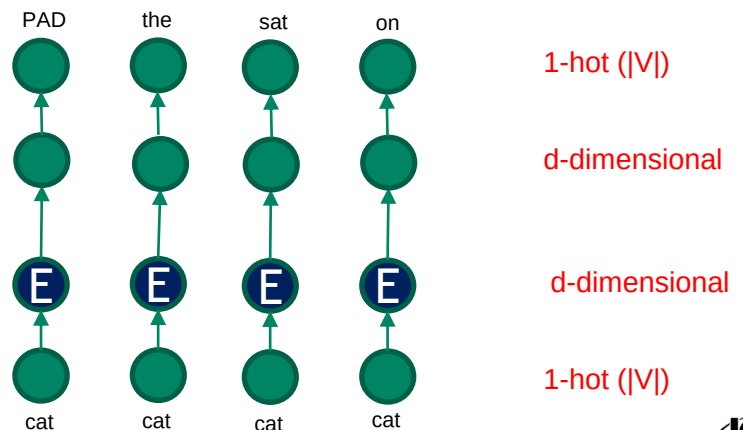
WORD2VEC

- Mikolov et al., 2013 (quando estava na Google)
- Modelo Linear (treina rapidamente)
- Dois modelos para treinamento de *embeddings* de forma não supervisionada :

Continuous Bag-of-Words (CBOW)



Skip-Gram



CONCLUSÕES

- A representação "clássica" de dados de texto é feita usando BoW e codificação 1-hot, o que pode levar a uma baixa precisão devido à dispersão e falta de generalização semântica
- O uso de representações distribuídas (também conhecidas como *embeddings* de palavras) adiciona algum conhecimento a priori à representação de entrada.
- *Embeddings* de palavras são normalmente usados como entrada para outras tarefas de PLN.