PROCESSAMENTO DE LINGUAGEM NATURAL

Embeddings

TÓPICOS

1. Introdução
2. Word2Vec
3. Doc2Vec

INTRODUÇÃO

- O conceito de *embeddings* se refere a uma representação em um espaço vetorial contínuo, de menor dimensionalidade aprendida, ou gerada a partir de uma representação de maior dimensionalidade
- As *embeddings* podem ser aprendidas/geradas para variáveis discretas (ex.: documentos, sentenças e palavras)
- Ao gerar uma representação de menor dimensionalidade, deve-se tentar preservar as características do espaço de maior dimensionalidade
- Ao gerar um espaço de menor dimensionalidade, o processamento por parte dos algoritmos de aprendizado de máquina torna-se mais rápido e consome menos espaço

INTRODUÇÃO

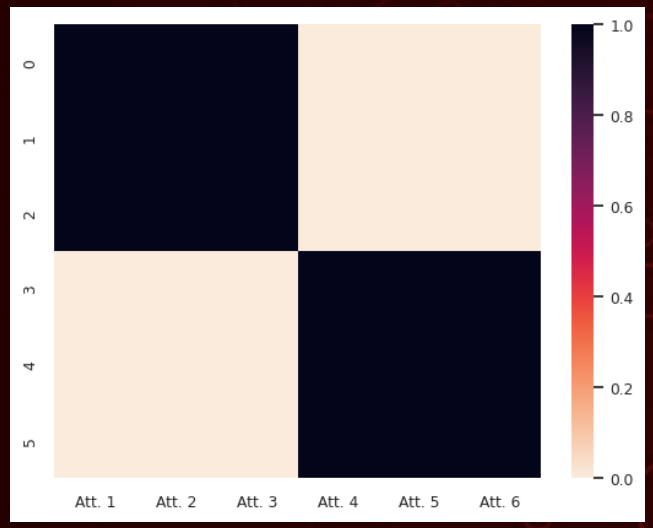
Espaço Original

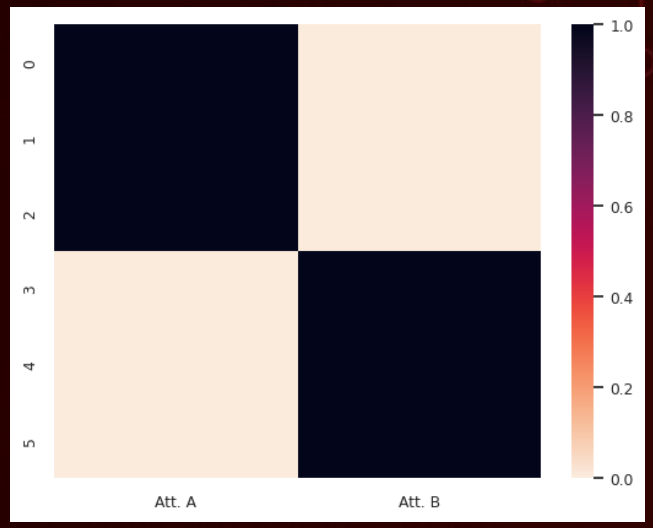
	Att. 1	Att. 2	Att. 3	Att. 4	Att. 5	Att. 6
0	1	1	1	0	0	0
1	1	1	1	0	0	0
2	1	1	1	0	0	0
3	0	0	0	1	1	1
4	0	0	0	1	1	1
5	0	0	0	1	1	1

Espaço *Embedding*

	Att. A	Att. B
0	1	0
1	1	0
2	1	0
3	0	1
4	0	1
5	0	1

Matriz de Similaridade





INTRODUÇÃO

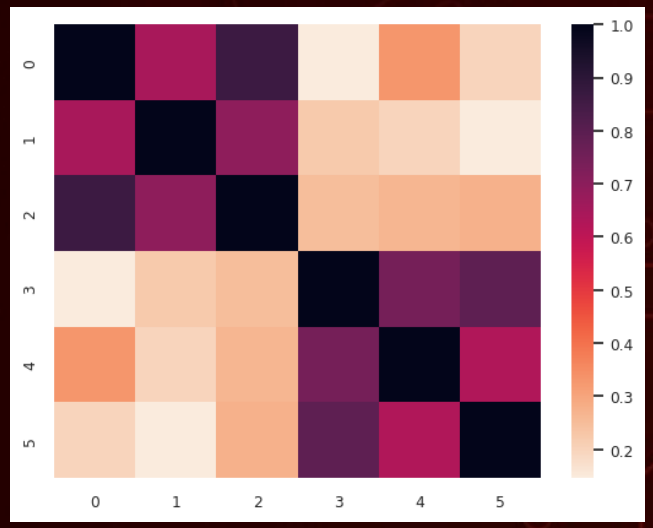
Espaço Original

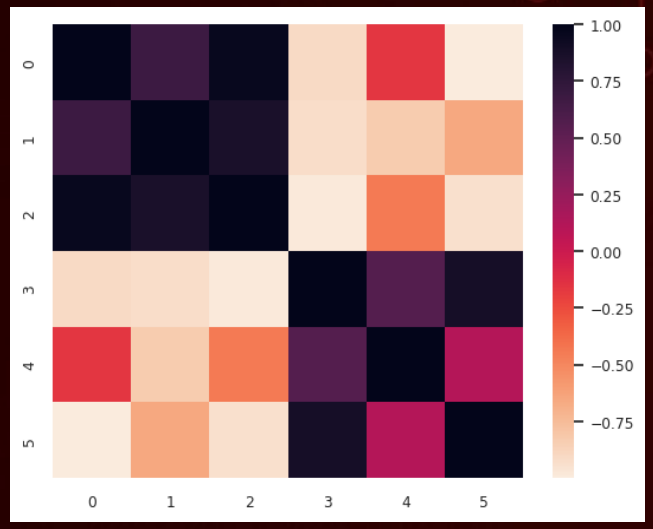
	Att. 1	Att. 2	Att. 3	Att. 4	Att. 5	Att. 6
0	3	2	1	0	1	0
1	1	3	4	0	0	1
2	3	1	2	1	0	0
3	0	1	0	4	1	3
4	1	0	0	1	2	3
5	0	0	1	5	4	1

Espaço *Embedding*

	Att. A	Att. B
0	-2.419284	0.824008
1	-3.188673	-1.697054
2	-2.209397	0.077006
3	2.705855	0.294023
4	1.105915	2.087878
5	4.005585	-1.585861

Matriz de Similaridade

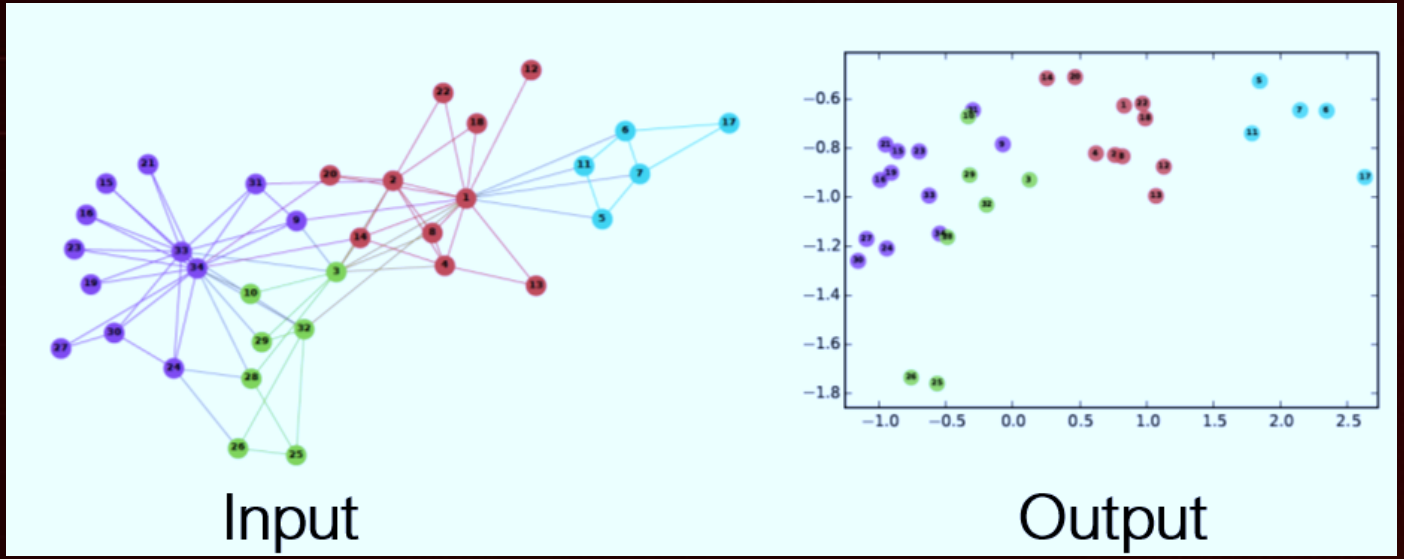




Correlação de Pearson entre as duas matrizes é 0,9575.

EMBEDDINGS

- **O uso de *embeddings*, além de ser particularmente útil para reduzir o número de dimensões, também possui uma série de outros efeitos interessantes, de acordo com o tipo do dado que estão representando**
  - **Representar grafos no espaço-vetorial → permite utilizar qualquer algoritmo baseado no modelo espaço vetorial**



EMBEDDINGS

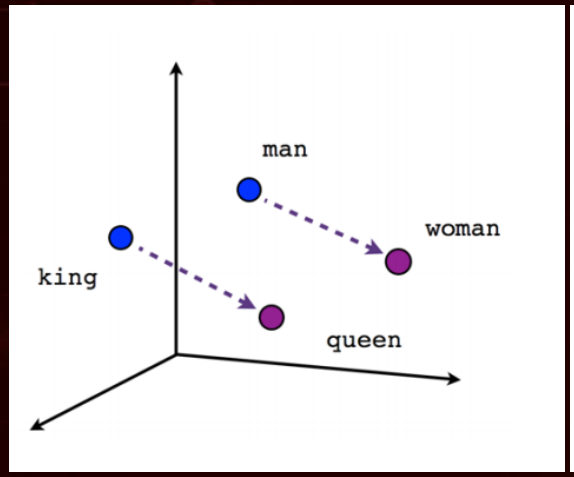
- **O uso de *embeddings*, além de ser particularmente útil para reduzir o número de dimensões, também possui uma série de outros efeitos interessantes, de acordo com o tipo do dado que estão representando**
  - Representar palavras (*word embeddings*) →
    - Permite capturar a semântica das palavras
    - Operações entre os vetores de palavras são significativas

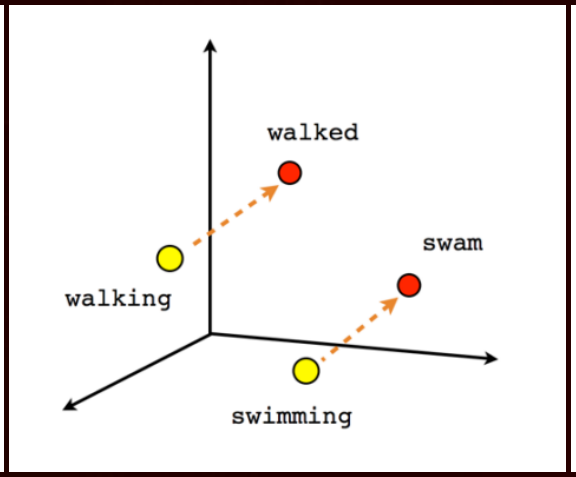
	King	Queen	Princess	Boy
Royal	0,99	0,99	0,99	0,01
Male	0,99	0,02	0,01	0,98
Female	0,02	0,99	0,99	0,01
Age	0,70	0,60	0,10	0,20

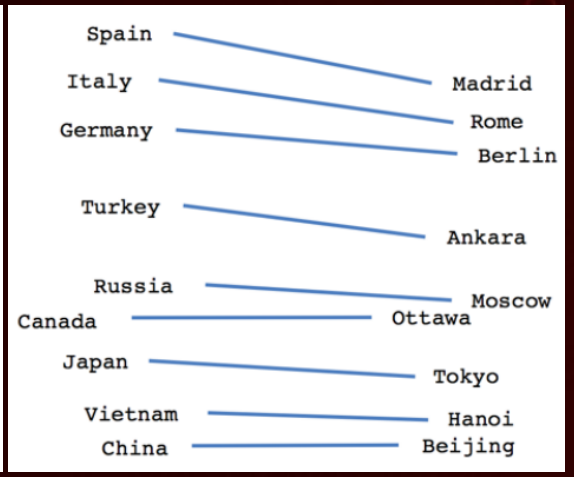
$$f(king) - f(queen) \approx f(man) - f(woman)$$

EMBEDDINGS

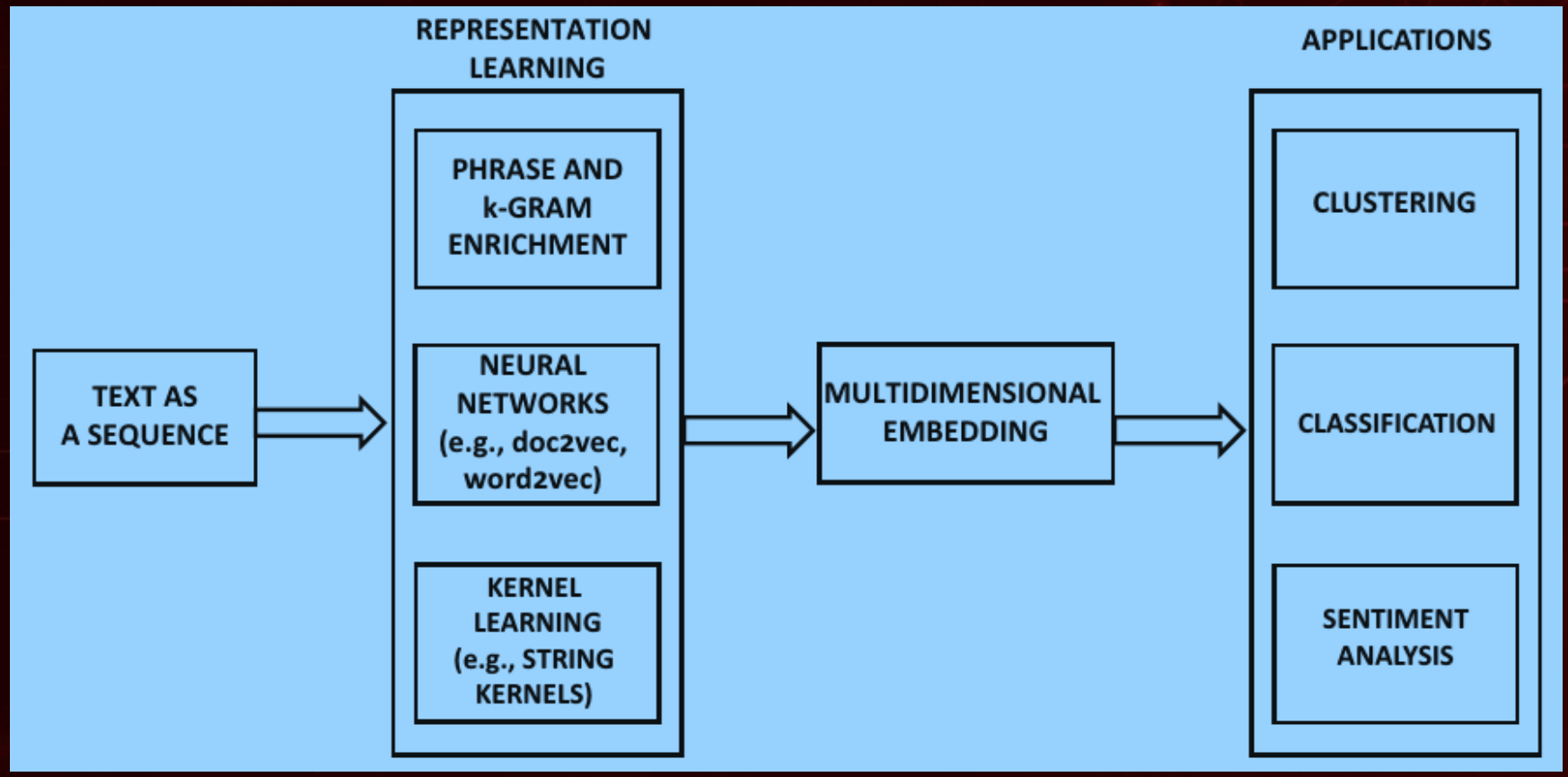
- **O uso de *embeddings*, além de ser particularmente útil para reduzir o número de dimensões, também possui uma série de outros efeitos interessantes, de acordo com o tipo do dado que estão representando**
  - Representar palavras (*word embeddings*) →
    - Permite capturar a semântica das palavras
    - Operações entre os vetores de palavras são significativas





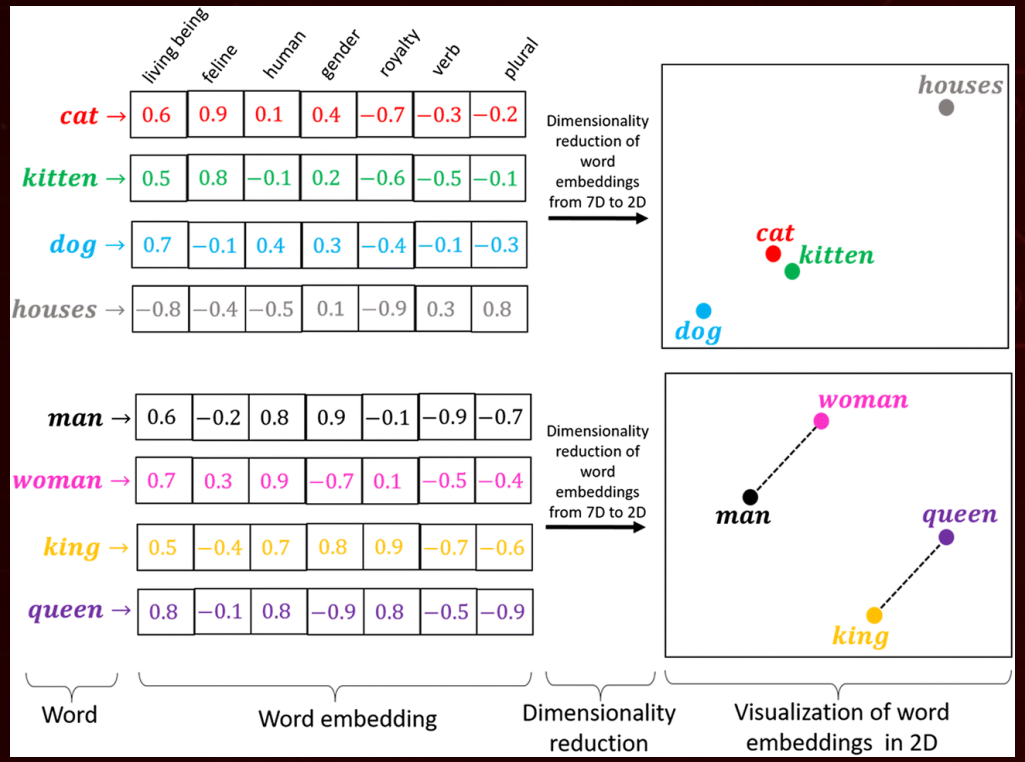


EMBEDDINGS


REPRESENTATION LEARNING
APPLICATIONS
TEXT AS A SEQUENCE
PHRASE AND k-GRAM ENRICHMENT
NEURAL NETWORKS (e.g., doc2vec, word2vec)
KERNEL LEARNING (e.g., STRING KERNELS)
MULTIDIMENSIONAL EMBEDDING
CLUSTERING
CLASSIFICATION
SENTIMENT ANALYSIS

WORD EMBEDDINGS

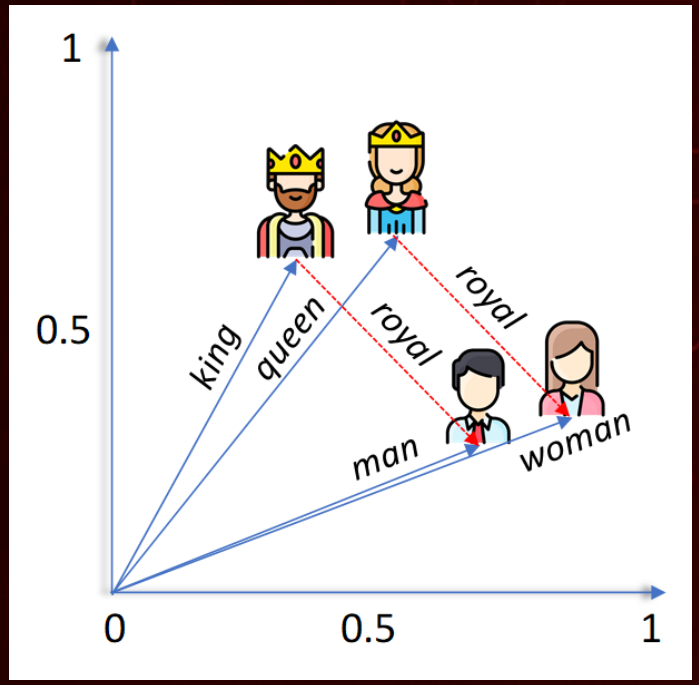
- Representações vetoriais densas (valores diferentes de zero)



- A dimensão é fixa (p. ex. 300)

WORD EMBEDDINGS

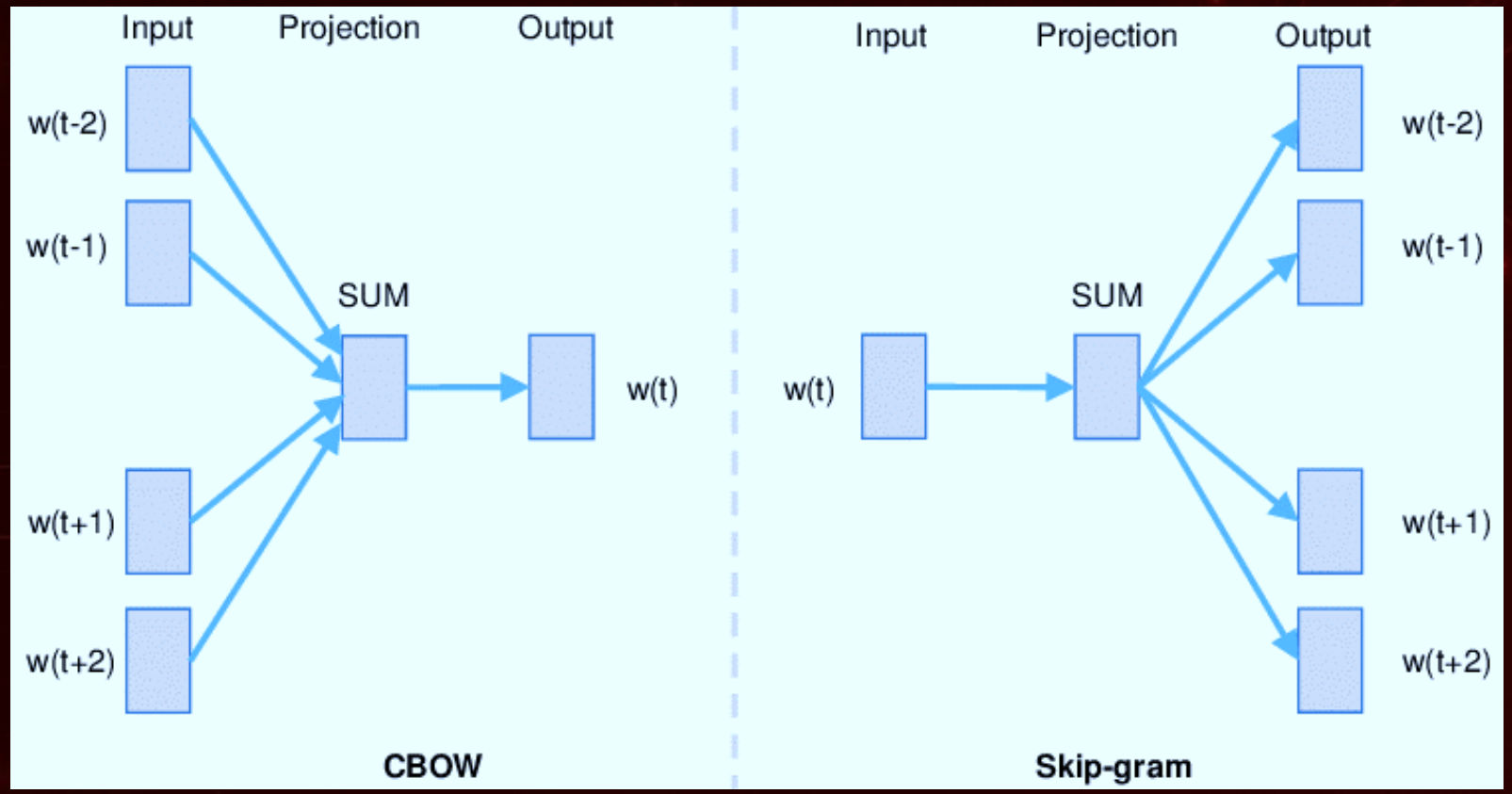
- v(king) + v(woman) - v(man) ~ v(queen)
- v(king) - v(royal) ~ v(man)
- v(queen) - v(royal) ~ v(woman)



Word2Vec

- A abordagem Word2Vec é uma das mais populares para geração de *word-embeddings* utilizando redes neurais
- Há duas variações da abordagem Word2Vec:
  - *Continuous Bag-Of-Words* (CBOW)
  - Skip-Gram

CBOW vs *Skip-Gram*



QUESTÕES IMPORTANTES

- **Número de dimensões**
  - Aumentar a dimensionalidade da *embedding* geralmente aumenta o poder discriminativo, porém, requer mais dados para obter melhores resultados
  - Em geral, o número de dimensões varia em algumas centenas
- **Tamanho do contexto: tipicamente varia entre 5 e 10**
- **Efeito de palavras frequentes e pouco discriminativas (ex.: “the”):**
  - Podem interferir nos resultados
  - Pode-se removê-las ou adotar um esquema de amostragem inversamente proporcional à frequência
- **Também é interessante a identificação de frases para a geração de *embeddings*, por exemplo, “Apple Store”**

FastText

- Extensão da abordagem Skip-Gram
- Ao invés de palavras como entrada, são consideradas sequências de *n* caracteres ($3 \leq n \leq 6$)
- Permite aprender uma estrutura interna das palavras
- Tende a obter melhores resultados em línguas morfologicamente ricas
- Uma vez obtida a *embedding* das sequências de *n*-caracteres, a *embedding* da palavra é obtida por meio da soma das *embeddings* das sequências de caracteres

Document Embeddings

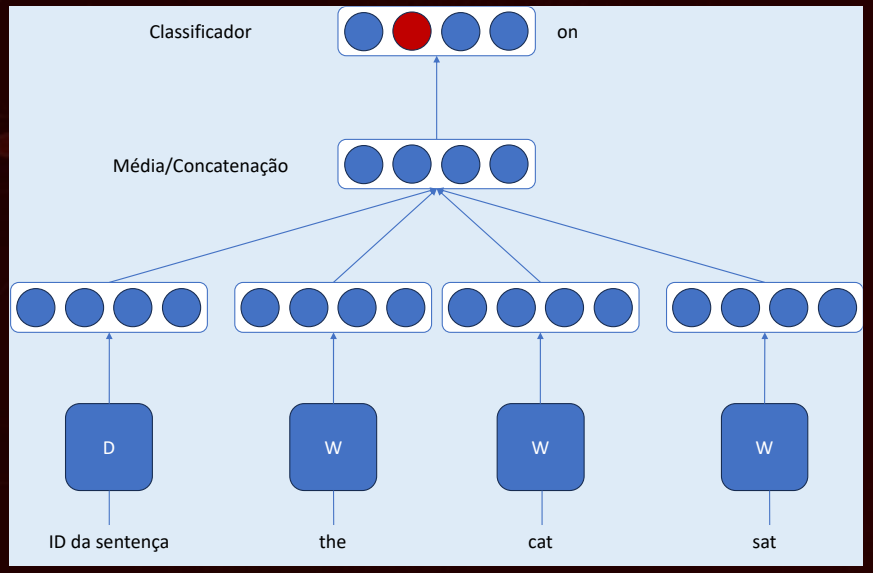
- **Tudo o que foi visto é útil, pois:**
  - **Podemos gerar as representações dos documentos a partir das *word embeddings***
  - **Utilizar abordagens parecidas à da geração das *word embeddings*, porém, para a geração das *embeddings* dos documentos**
- **As abordagens apresentadas na seção anterior, apesar dos diversos aspectos interessantes que elas apresentam, para um dos propósitos dessa aula, é necessária a representação dos documentos, e não das palavras**

Doc2Vec

- A partir das *word embeddings*, pode-se obter os vetores das palavras do documento e fazer operações sobre esses vetores, tais como: Soma, Média, Soma ponderada, Valores mínimos ou máximos de cada dimensão
- Conhecida como *Paragraph Vectors* ou Doc2Vec
- Le e Mikolov (2014) propõem duas abordagens para gerar *embeddings* de trechos de texto, como parágrafos, ou documento, baseadas no Word2Vec
- As duas abordagens são:
  - *Distributed Memory* (PV-DM)
  - *Distributed Bag-of-Words* (PV-DBOW)

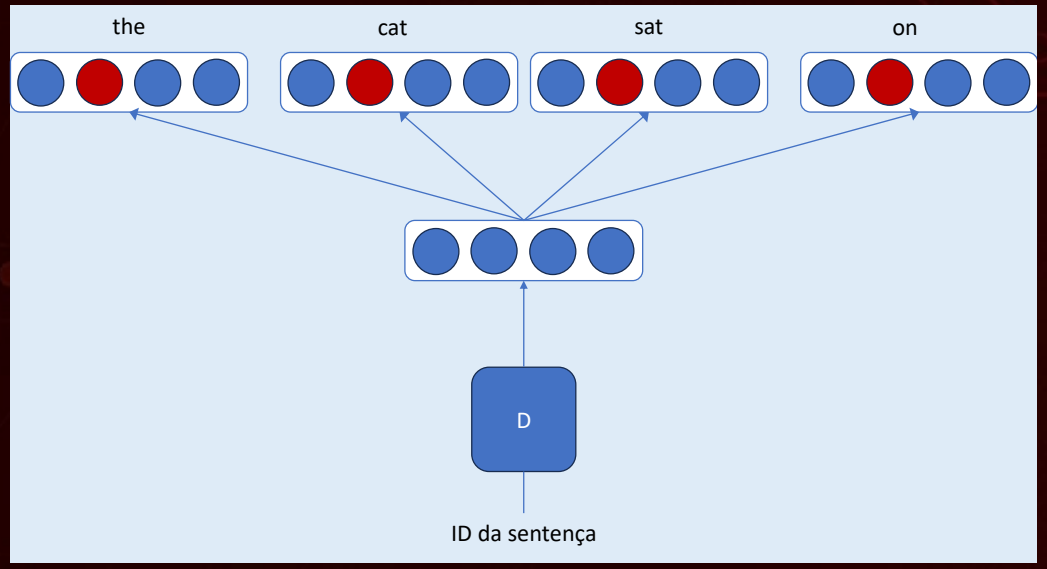
PV-DM

- São consideradas como entradas os vetores *one-hot* de cada parágrafo e das palavras do contexto
- Portanto, a entrada do PV-DM é similar a do CBOW com a adição dos vetores *one-hot* dos parágrafos
- Uma outra diferença em relação ao CBOW é que o objetivo é prever a próxima palavra do contexto



PV-DBOW

- No modelo PV-DBOW, o objetivo é prever as palavras do contexto dado o parágrafo como entrada
- Esse modelo é análogo ao modelo Skip-Gram utilizado na geração das *word embeddings*



PRÁTICA

- Nas implementações para a geração das *word* e *document embeddings*, vamos utilizar:
  - a biblioteca Gensim: https://radimrehurek.com/gensim/
  - bibliotecas auxiliares para a manipulação de dados e pré-processamento dos textos
- Vamos aprender:
  - a geração de *word embeddings* para uma base
  - a utilização de *word embeddings* já treinadas

O QUE VIMOS?

- Introdução
- Word2Vec
- Doc2Vec

PRÓXIMA VIDEOAULA

- **Prática: *Embeddings***

REFERÊNCIAS

- Curso de Tópicos em Inteligência Artificial
  - Prof. Rafael G. Rossi (UFMS)
  - https://www.youtube.com/@RafaelRossiTech/playlists

- Curso de Processamento de Linguagem Natural
  - Profa. Helena Caseli (UFSCar)