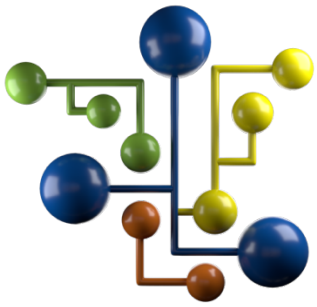

Data Science Academy

www.datascienceacademy.com.br

Machine Learning

Estudo de Caso - Buscador de Palavras em Texto Por Similaridade

Trabalhar com palavras é sempre um desafio. Mesmo para um pequeno corpus, sua rede neural (ou qualquer tipo de modelo) precisa suportar milhares de entradas e saídas discretas.

Além das palavras numéricas brutas, a técnica padrão de representar palavras como vetores one-hot (por exemplo, "um" = [0 0 0 1 0 0 0 0 ...]) não captura nenhuma informação sobre relacionamentos entre palavras.

Os vetores de palavras (word embeddings) solucionam esse problema, representando palavras em um espaço vetorial multidimensional. Isso pode levar a dimensionalidade do problema de centenas de milhares para apenas centenas. Além disso, o espaço vetorial é capaz de capturar relações semânticas entre as palavras em termos de distância e aritmética vetorial.

Existem algumas técnicas para criar vetores de palavras. O algoritmo Word2vec prevê palavras em um contexto (por exemplo, qual é a palavra mais provável na frase "O gato _____ no telhado"), enquanto os vetores GloVe são baseados em contagens globais em todo o corpus.

Usaremos o GloVe agora neste estudo de caso e o Word2vec na próxima aula.

Modelo GloVe

O GloVe é uma técnica de vetor de palavras (embeddings). Os vetores de palavras colocam as palavras em um espaço vetorial, onde palavras semelhantes se agrupam e palavras diferentes se repelem. A vantagem do GloVe é que, diferentemente do Word2vec, o GloVe não depende apenas de estatísticas locais (informações de contexto local das palavras), mas incorpora estatísticas globais (co-ocorrência de palavras) para obter vetores de palavras. Mas há bastante sinergia entre o GloVe e o Word2vec.

E não se surpreenda ao saber que a ideia de usar estatísticas globais para derivar relacionamentos semânticos entre palavras remonta a um longo caminho. GloVe significa "Global Vectors" ou “Vetores Globais". E, como mencionado anteriormente, o GloVe captura estatísticas globais e estatísticas locais de um corpus, a fim de criar vetores de palavras. Mas precisamos de estatísticas globais e locais?

Acontece que cada tipo de estatística tem sua própria vantagem. Por exemplo, o Word2vec, que captura estatísticas locais, se sai muito bem em tarefas de analogia. No entanto, um método como o LSA (Latent Semantic Analysis), que usa apenas estatísticas globais, não funciona bem em tarefas de analogia.

Como Funciona o GloVe

Dado um corpus com V palavras, a matriz de co-ocorrência X será uma matriz V x V, onde a i-linha e a j-ésima coluna de X, X_ij indica quantas vezes a palavra i co-ocorreu com a palavra j. Um exemplo de matriz de co-ocorrência pode ter a seguinte aparência.

	the	cat	sat	on	mat
the	0	1	0	1	1
cat	1	0	1	0	0
sat	0	1	0	1	0
on	1	0	1	0	0
mat	1	0	0	0	0

Como obtemos uma métrica que mede a similaridade semântica entre as palavras? Para isso, você precisará de três palavras por vez. Deixe-me apresentar concretamente esta afirmação.

Probability and Ratio	$k = solid$	$k = gas$	$k = water$	$k = fashion$
$P(k\|ice)$	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
$P(k\|steam)$	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
$P(k\|ice)/P(k\|steam)$	8.9	8.5×10^{-2}	1.36	0.96

Onde:

P_ik / P_jk em que P_ik = X_ik / X_i

Aqui P_ik denota a probabilidade de ver as palavras i e k juntas, o que é calculado dividindo o número de vezes que i e k apareceram juntas (X_ik) pelo número total de vezes que as palavras apareciam no corpus (X_i).

Você pode ver que, dadas duas palavras, ou seja, *ice* (gelo) e *steam* (vapor), se a terceira palavra k é muito semelhante a *ice*, mas irrelevante a *steam* (por exemplo, k = *fashion*), P_ik / P_jk será muito alto (> 1), e muito semelhante a *steam*, mas irrelevante para *ice* (por exemplo,

k = *gas*), P_ik / P_jk será muito pequeno (<1), e se estiver relacionado ou não relacionado a nenhuma das palavras, o P_ik / P_jk estará próximo de 1.

Portanto, se pudermos encontrar uma maneira de incorporar P_ik / P_jk na computação de vetores de palavras, alcançaremos o objetivo de usar estatísticas globais ao aprender vetores de palavras.

E é exatamente isso que faz o modelo GloVe. No link abaixo você encontra o paper original:

https://nlp.stanford.edu/pubs/glove.pdf

E ao final do capítulo você encontra o Jupyter Notebook "**07-DSA-Cap12-GloVe.ipynb**" com o modelo completo. Leia atentamente cada célula, execute o notebook e experimente o GloVe.

Boa aula.