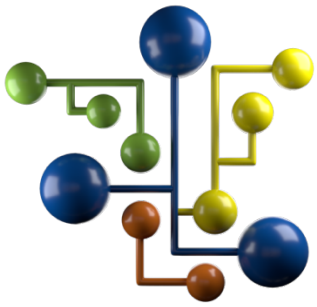

Data Science Academy

www.datascienceacademy.com.br

Machine Learning

Estudo de Caso - Previsão de Palavras com Base no Contexto e Visualização com PCA

O Word2vec é uma rede neural de duas camadas que processa o texto “vetorizando” as palavras. Sua entrada é um corpus de texto e sua saída é um conjunto de vetores; vetores de recursos que representam palavras nesse corpus. Embora o Word2vec não seja uma rede neural profunda, ele transforma o texto em uma forma numérica que as redes neurais profundas podem entender.

O Word2vec é um método de calcular representações vetoriais de palavras e foi desenvolvido por uma equipe de pesquisadores do Google liderada por Tomas Mikolov. O Google hospeda uma versão de código aberto do Word2vec lançada sob uma licença Apache 2.0. Em 2014, Mikolov deixou o Google para o Facebook e, em maio de 2015, foi concedida ao Google uma patente para o método, que não revoga a licença do Apache sob a qual foi lançado. Aqui o paper original:

https://arxiv.org/pdf/1301.3781.pdf

As aplicações do Word2vec vão além da análise de sentenças. Também pode ser aplicado a genes, códigos, curtidas em redes sociais, listas de reprodução, gráficos de mídias sociais e outras séries verbais ou simbólicas nas quais os padrões podem ser discernidos.

Por quê? Como as palavras são simplesmente estados discretos, como os outros dados mencionados acima, estamos simplesmente procurando as probabilidades de transição entre esses estados, ou seja, a probabilidade de que elas co-ocorram. Assim, gene2vec, like2vec e follower2vec são todos possíveis.

O objetivo e a utilidade do Word2vec é agrupar os vetores de palavras semelhantes no espaço de vetores. Ou seja, ele detecta semelhanças matematicamente. O Word2vec cria vetores que são representações numéricas distribuídas de recursos de palavras, recursos como o contexto de palavras individuais. Faz isso sem intervenção humana.

Com dados e contextos suficientes, o Word2vec pode fazer suposições altamente precisas sobre o significado de uma palavra com base no contexto. Essas suposições podem ser usadas para estabelecer a associação de uma palavra com outras palavras (por exemplo, "homem" é "garoto" e "mulher" é "garota") ou agrupar documentos e classificá-los por tópico. Esses agrupamentos podem formar a base da pesquisa, análise de sentimentos e recomendações em diversos campos, como pesquisa científica, mineração de documentos legais e jurídicos, comércio eletrônico e gerenciamento de relacionamento com o cliente.

A saída da rede neural do Word2vec é um vocabulário no qual cada item tem um vetor anexado, que pode ser alimentado em uma rede de aprendizado profundo ou simplesmente consultado para detectar relações entre palavras.

Medindo a similaridade do cosseno, nenhuma similaridade seria expressa como um ângulo de 90 graus, enquanto a similaridade total de 1 é um ângulo de 0 graus, ou seja, a palavra Suécia é igual à palavra Suécia, enquanto a palavra Noruega tem uma distância de cosseno de 0,760124 da palavra Suécia, por exemplo.

Similaridade de Cosseno

Entre diferentes métricas de distância, a similaridade de cosseno é mais intuitiva e mais usada no Word2vec. É um produto normalizado de 2 vetores e essa relação define o ângulo entre eles. Dois vetores com a mesma orientação têm uma similaridade de cosseno de 1, dois vetores a 90 ° têm uma similaridade de 0 e dois vetores diametralmente opostos têm uma similaridade de -1, independentemente de sua magnitude.

Uma vez que as palavras são representadas por vetores, a tarefa de encontrar palavras semelhantes ou diferentes torna-se mais fácil. Quaisquer combinações de vetores resultam em um novo vetor e as distâncias do cosseno ou outras medidas de similaridade podem ser usadas. É assim que resolvemos a famosa equação que define o Word2vec:

'rei - homem + mulher = rainha'

Com um modelo Word2vec, conseguimos associar palavras com base no seu contexto, usando nossa boa e velha Matemática.

Isso é o que faremos neste Estudo de Caso, que você encontra no Jupyter Notebook "**08-DSA-Cap12-Word2vec**". Leia os comentários com atenção, execute as células, modifique os exemplos e compreenda o funcionamento do Word2vec.

Boa aula.