Trabalho Final CCD2

Aplicação de alg de predição + rede neural com embedding + BERTimbau

Conteúdos

01

BoW e TF-IDF

02

NB e KNN

03

Word2vec

04

BERTimbau

BoW - Bag of Words

O Bag of Words é uma representação de texto que transforma um documento em um conjunto de palavras, ignorando a ordem e a estrutura gramatical. Cada documento é tratado como um “saco” de palavras, onde a presença ou ausência de palavras específicas é registrada.

“John gosta de assistir filmes.
Mary também gosta de
filmes.”

[\text{“John”: 1, “gosta”: 2, “de”: 1, “assistir”: 1, “filmes”: 2, “Mary”: 1, “também”: 1}]

TF-IDF Term Frequency-Inverse Document Frequency

Frequency (TF): Mede a frequência de um termo específico em um documento.

Inverse Document Frequency (IDF):Mede a importância do termo em todo o corpus.

TF-IDF: Combina TF e IDF para calcular a relevância de um termo em um documento específico.

"O gato gosta de peixe."

"O cachorro gosta de osso."

TF Frase 1 "gosta" aparece 1 vez em 5 palavras totais. Então, TF(gosta,Frase 1)=1/5=0.2

TF Frase 2: "gosta" aparece 1 vez em 5 palavras totais. Então, TF(gosta,Frase 2)=1/5=0.2

IDF
"gosta" aparece em ambos os documentos. Então, IDF(gosta)=log(2/2)=log(1)=0

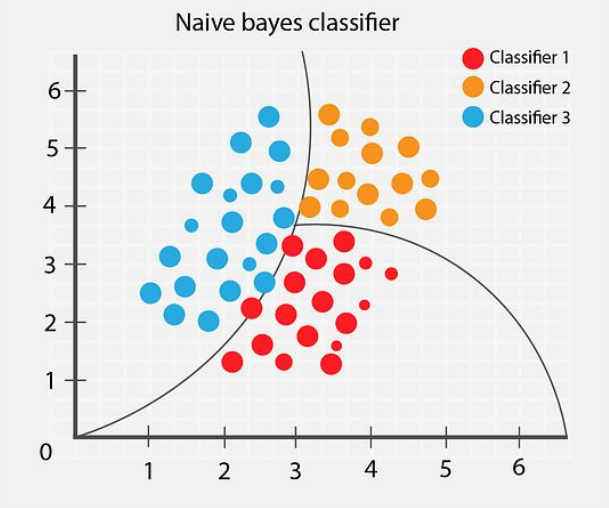
TF-IDF 1 (gosta,Frase 1) =0.2×0=0

TF-IDF 2(gosta,Frase 2) =0.2×0=0

Naive Bayes - NB

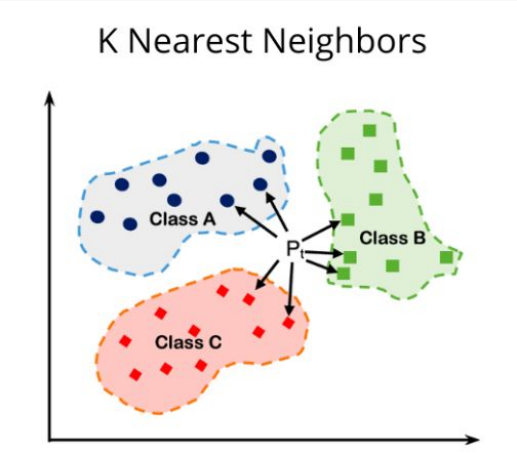
O Teorema de Bayes descreve a probabilidade de um evento, baseado em conhecimento prévio de condições que possam estar relacionadas ao evento.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$



K-Nearest Neighbors - KNN

é um algoritmo baseado em instâncias que classifica um ponto de dados com base na proximidade dos pontos de dados vizinhos. Ele é chamado de “preguiçoso” porque não faz suposições sobre a distribuição dos dados e não constrói um modelo explícito durante a fase de treinamento. Em vez disso, ele armazena todos os casos de treinamento e faz a classificação apenas quando um novo ponto de dados precisa ser classificado



Word2vec

Vetores independentes de contexto.

Redes neurais simples (CBOW e Skip-gram).

Tarefas simples de PLN, como similaridade de palavras.

vs

BERTimbau

Vetores dependentes de contexto.

Transformadores bidirecionais.

Tarefas complexas de PLN, como análise de sentimentos, tradução automática e resposta a perguntas.


A picture is worth a thousand words