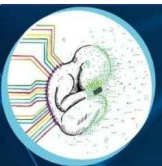




Análise de sentimentos com Google Colaboratory

Michel Ferreira Batista

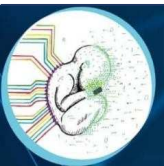




17ª SEMANA NACIONAL DE
CIÊNCIA E TECNOLOGIA
Inteligência Artificial: A Nova Fronteira da Ciência Brasileira

Análise de sentimentos

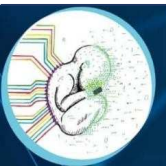




17ª SEMANA NACIONAL DE
CIÊNCIA E TECNOLOGIA
Inteligência Artificial: A Nova Fronteira da Ciência Brasileira

O que é análise de sentimentos?

“Análise de Sentimentos ou Mineração de opiniões refere-se ao uso de PLN com o objetivo de identificar, extrair e quantificar a polaridade expressamente dos dados coletados.”

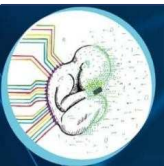


O que é análise de sentimentos?

É um campo dentro do Processamento de Linguagem Natural (PLN) que constrói sistemas que identificam e extraem opiniões dentro de textos.

Normalmente, além de identificar a opinião, esses sistemas extraem atributos da expressão, como:

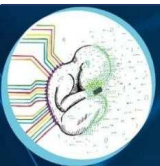
- Assunto: o que está sendo falado;
- Detentor de opinião: quem é a pessoa ou entidade que expressa a opinião;
- Polaridade: se o falante expressou uma opinião positiva ou negativa.



O que é análise de sentimentos?

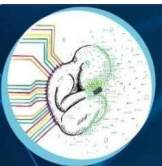
- AS está em grande expansão
- Pode ser a porta de entrada para outras análises;
- Informações públicas e privadas na Internet estão sempre crescendo

Assim, existem cada vez mais textos que expressam opiniões em sites de reviews sobre produtos, fóruns, blogs e mídias sociais.



O que é análise de sentimentos?





17ª SEMANA NACIONAL DE
CIÊNCIA E TECNOLOGIA
Inteligência Artificial: A Nova Fronteira da Ciência Brasileira


O que é análise de sentimentos?

Ótimo



24/10/2020

Adorei! Ótima ma compra e preço b om, lindo designer

Simone  comprei e avaliei

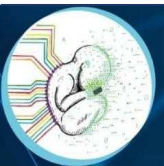
Smartphone Samsung Galaxy A10s



24/10/2020

No geral o produto é bom. O touch não é tão sensível como nos demais aparelhos de celular que eu já tive. O cabo USB do carregador muito curto. Tive que comprar um outro à parte.

Eduardo  comprei e avaliei



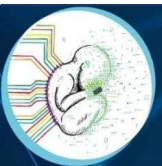
Qual a importância análise de sentimentos?

Estima-se que 80% dos dados do mundo estejam desestruturados e desorganizados.

A maior parte deles são provindos de textos, como e-mails, bate-papos, mídias sociais, pesquisas, artigos e documentos.

Estes textos geralmente são difíceis, demorados e caros para analisar, entender e classificar.

Os sistemas de Análise de Sentimento permitem que as empresas entendam esse mar de textos não-estruturados.

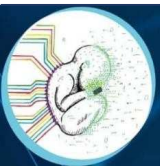


17ª SEMANA NACIONAL DE
CIÊNCIA E TECNOLOGIA
Inteligência Artificial: A Nova Fronteira da Ciência Brasileira

Qual a importância análise de sentimentos?

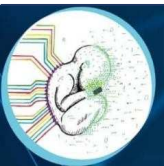
Imagine como seria classificar manualmente milhares de tweets, conversas de suporte ao cliente ou comentários sobre um produto no Facebook?

A análise de sentimentos permite entender estes dados em escala de maneira eficiente e econômica. Isto é apenas um pouco do que a Análise de Sentimento é capaz de fazer.



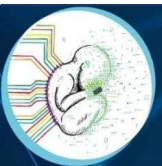
Experiência Acadêmica

- Dificuldades da AS:
 - Negações: Uso de negações altera a polaridade de uma sentença ou expressão:
 - Exemplo: “Eu gosto do meu Smartphone.”
 - “Eu não gosto do meu Smartphone.”
 - Metáforas, tons de Ironias.
 - Mais de um sentimento em uma sentença:
 - Exemplo: “Eu gosto desse smartphone, mas se pudesse compraria um melhor”
 - ???????



• Dicionários Léxicos

- Base de dados com palavras associadas a sentimentos;
- Negativo: -1 Neutro: 0 Positivo: +1
- Baseados em áreas de conhecimento:
 - Saúde, Tecnologia, Psicologia,
- Vantagem:
 - Não precisa ter dados de treinamento, ou seja, rotular;
- Desvantagem:
 - Depender do Idioma e manutenção da base;



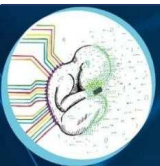
Abordagens da AS

- **Machine Learning (Aprendizado de Máquina)**
 - Modelar dados de texto na forma de Bag of Words (saco de palavras)
 - Vantagem:
 - Não depende de Idioma, tudo depende do treinamento;
 - Desvantagem:
 - Necessita de dados de treinamento;

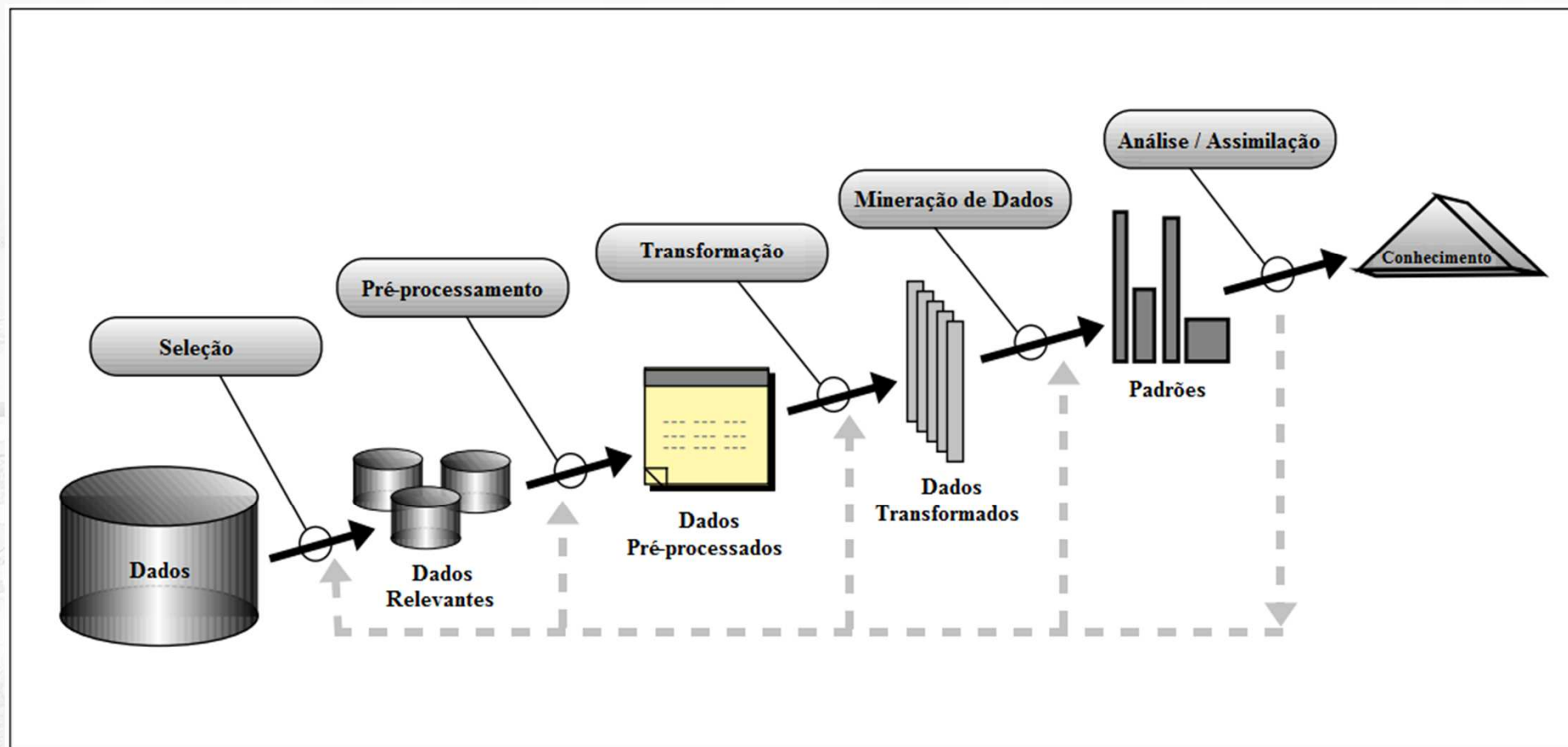


Mineração de Texto



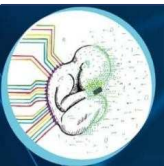


Mineração de Texto



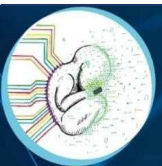


Bag of Words



Bag of words

Segundo Maalej et al. (2016) a técnica de saco de palavras (Bag of words) é uma das técnicas mais utilizadas para extração de recursos em textos. Baseada em uma medida estatística, TF (Term Frequency), que tem o intuito de indicar a importância de um termo de um documento em relação a corpus linguístico por exemplo, de acordo com a frequência de termos.



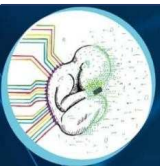
Bag of words

Sejam dadas duas frases

- Por favor remova os anúncios do aplicativo. O aplicativo é bom, mas com tanto anúncio fica irritante.
- Ter comida postado em anúncios no aplicativo perda de peso não é útil.

anúncios, aplicativo, bom, com, comida, de, do, em, favor, fica, irritante, mas, no, não, os, perda, peso, por, postado, remova, tantos, ter, útil

por: 17, favor: 8, remova: 19, os: 14, anúncios: 0, do: 6, aplicativo: 1, bom: 2, mas: 11, com: 3, tantos: 20, fica: 9, irritante: 10, ter: 21, comida: 4, postado: 18, em: 7, no: 12, perda: 15, de: 5, peso: 16, não: 13, útil: 22



Bag of words

Vetor de frequência de cada frase

Por favor remova os anúncios do aplicativo. O aplicativo é bom, mas com tantos anúncios fica irritante.

(0, 17) 1
(0, 8) 1
(0, 19) 1
(0, 14) 1
(0, 0) 2
(0, 6) 1
(0, 1) 2
(0, 2) 1
(0, 11) 1
(0, 3) 1
(0, 20) 1
(0, 9) 1
(0, 10) 1

Vetor: [2 2 1 1 0 0 1 0 1 1 1 1 0
0 1 0 0 1 0 1 1 0 0]

Ter comida postado em anúncios no aplicativo perda de peso não é útil

(1, 0) 1
(1, 1) 1
(1, 21) 1
(1, 4) 1
(1, 18) 1
(1, 7) 1
(1, 12) 1
(1, 15) 1
(1, 5) 1
(1, 16) 1
(1, 13) 1
(1, 22) 1

Vetor: [1 1 0 0 1 1 0 1 0 0 0 0 1
1 0 1 1 0 1 0 0 1 1]



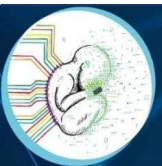
Classificador Naives Bayes



Classificador Naives Bayes

Uma das técnicas de mineração de dados amplamente utilizada é a classificação de dados. A classificação consiste no processo de encontrar, através de aprendizado de máquina, um modelo ou função que descreva diferentes classes de dados [Han e Kamber 2006].





17ª SEMANA NACIONAL DE
CIÊNCIA E TECNOLOGIA

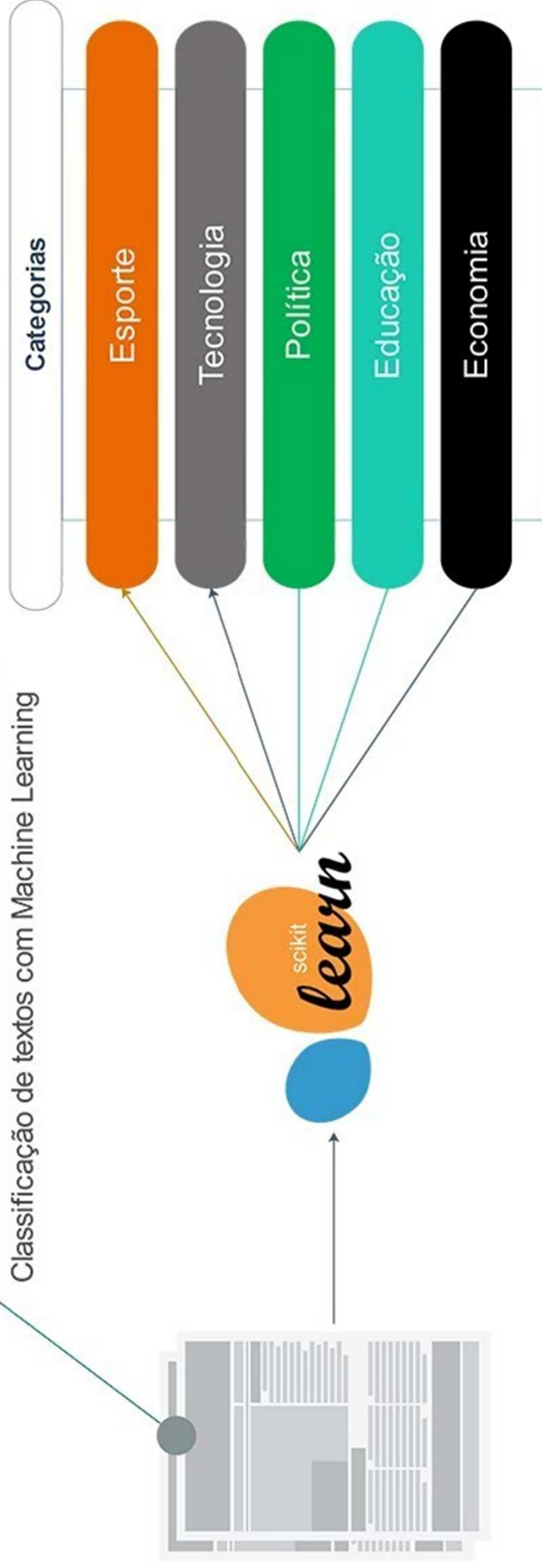
Inteligência Artificial: A Nova Fronteira da Ciência Brasileira

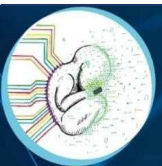
O objetivo da classificação é rotular, automaticamente, novas instâncias da base de dados com uma determinada classe aplicando o modelo ou função “aprendidos”. Este modelo é baseado no valor dos atributos das instâncias de treinamento.



Machine Learning

Classificação de textos com Machine Learning





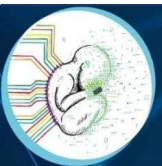
EM QUAL CLASSIFICAÇÃO CADA NOTÍCIA MELHOR REPRESENTA?

“Após reunião nesta quarta (8), no Recife, reitores anunciaram ações de conscientização sobre importância do trabalho acadêmico para a população.”

“O Supremo Tribunal Federal (STF) decidiu que não é preciso o aval do Legislativo para privatizar subsidiárias de estatais e controladas.”

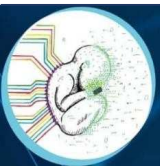
“Colombiano marca duas vezes, Galo vira no Pacaembu e elimina o Santos da Copa do Brasil”

“O presidente Jair Bolsonaro avançou dois sinais, ultrapassou pelo acostamento e passou a mais de 100 por um pardal para chegar mais rápido ao Congresso e levar sua proposta de mudanças no Código de Trânsito Brasileiro.”

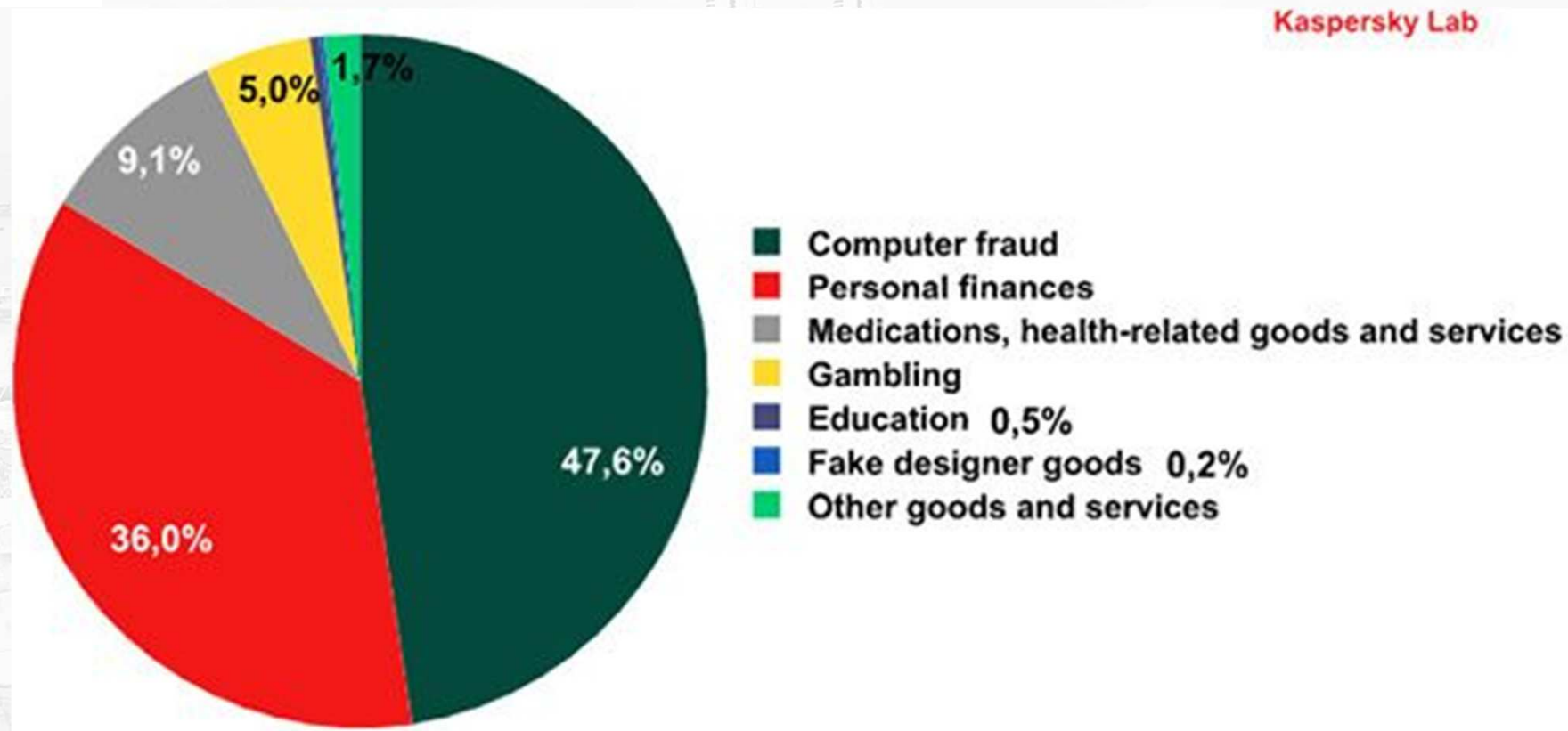


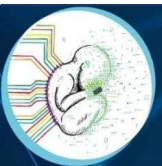
A classificação pode ser especializada na categorização textual, que consiste na organização de documentos em tópicos pré estabelecidos. Esta categorização tem diversas aplicações na área de Recuperação de Informação, tais como detecção de SPAM, organização automática de e-mails, identificação de páginas com conteúdo adulto e detecção de expressões multipalavras [Manning et al. 2008].





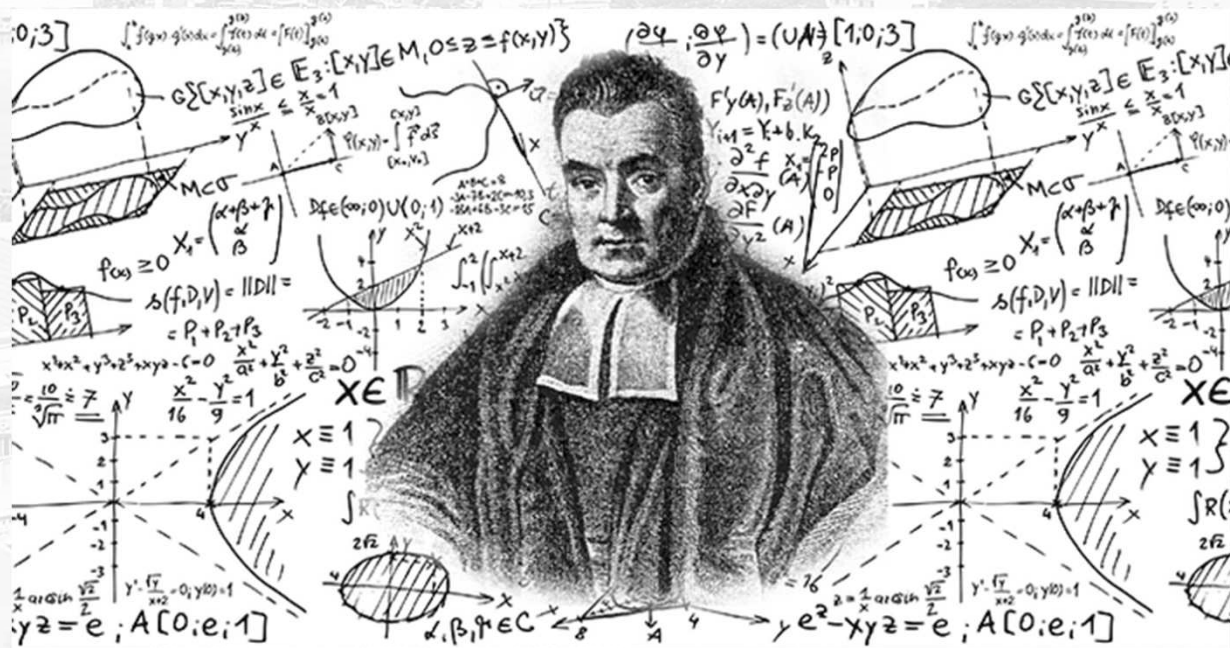
Classificação do spam em Setembro de 2012

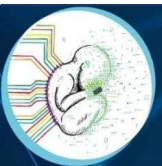




Naives Bayes

O algoritmo “Naive Bayes” é um classificador probabilístico baseado no “Teorema de Bayes”, o qual foi criado por Thomas Bayes (1701 - 1761).

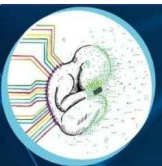




Naives Bayes

A principal característica do algoritmo, e também o motivo de receber “naive” (ingênuo) no nome, é que ele desconsidera completamente a correlação entre as variáveis (features). Ou seja, se determinada fruta é considerada uma “Maçã” se ela for “Vermelha”, “Redonda” e possui “aproximadamente 10cm de diâmetro”, o algoritmo não vai levar em consideração a correlação entre esses fatores, tratando cada um de forma independente.



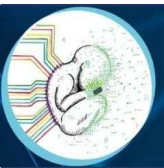


Naives Bayes

Silva (2016) afirma que o Teorema de Bayes é uma ferramenta da estatística, esse teorema é uma fórmula matemática utilizada para cálculos de probabilidades condicionais, descrevendo a probabilidade de um evento com base no conhecimento prévio das condições que podem estar relacionadas com o evento:

$$P(c | x) = \frac{P(X | c) P(c)}{P(x)}$$

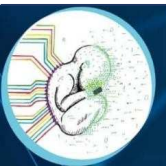
$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$



$$P(c | x) = \frac{P(X | c) P(c)}{P(x)}$$

$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$

Onde: $P(c | x)$ é a probabilidade posterior da classe alvo, $P(c)$ é a probabilidade original da classe, $P(X | c)$ é a possibilidade de que a probabilidade da classe preditora seja dada, $P(x)$ é a probabilidade original do preditor.



Probabilidade

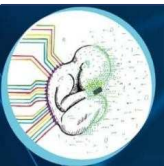
A probabilidade condicional trata da probabilidade de ocorrer um evento A , tendo ocorrido um evento B , ambos do espaço amostral S , ou seja, ela é calculada sobre o evento B e não em função o espaço amostral S .

A probabilidade de ocorrência de um evento A em relação a um evento ocorrido B é expressa como:

$$P(A|B)$$

Ex: $P(\text{cárie}|\text{dor de dente}) = 0.5$



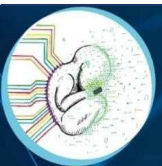


Naives Bayes

Naïve Bayes computa a probabilidade $P(c|x)$ de um documento pertencer a uma determinada classe a partir da probabilidade a priori $P(c)$ de um documento ser desta classe e das probabilidades condicionais $P(x_i|c)$ de cada termo x_i ocorrer em um documento da mesma classe.

$$P(c | x) = \frac{P(X | c) P(c)}{P(x)}$$

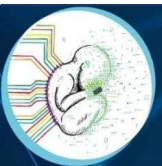
$$P(c | X) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c) \times P(c)$$



Naives Bayes

APRENDIZADO DE MÁQUINA

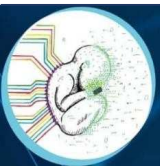
classe	d	Termos contidos na revisão do usuário
Desing	1	A tela não possui opção de inserir foto.
	2	A cor da tela não é boa, irrita a visão.
	3	O layout da tela é desorganizado, não sei buscar informação.
Usabilidade	4	Não consigo salvar meus exercícios.
	5	Gostaria de compartilhar minhas atividades na rede social.
Requisitos Funcionais	6	O botão salvar so aparece na barra de menu.
	7	gostaria de uma opção de monitor cardíaco



Naives Bayes

APRENDIZADO DE MÁQUINA

	TERMOS PRÉ PROCESSADOS				
	classe	d	Termos contidos na revisão do usuário	Termos por classes	conjunto treinamento
C	Desing	1	tela não possui opção inserir foto	19	38
		2	cor tela não boa irrita visão		
		3	layout tela desorganizado não sei buscar informação		
	Usabilidade	4	Não consigo salvar meus exercicios	11	
		5	Gostaria compartilhar minhas atividades redes sociais		
	Requisitos Funcionais	6	botão salvar so aparece barra menu	10	
		7	gostaria opção monitor cardíaco		
	?	8	falta tela cores identificar menus		



Naives Bayes

PROBABILIDADE À PRIORI

CALCULA A RAZÃO DO NÚMERO DE CLASSES DE ACORDO COM SUA CLASSIFICAÇÃO.

$$P(c) = \frac{c}{c(t)}$$

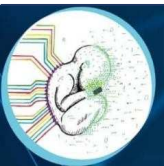
X

1º - P(c) - Probabilidade á priori de cada classe

P(Desing) = 3/7 | P(Usabilidade) = 2/7 | P(Requisitos) = 2/7

P(c)	PROB. A PRIORE
P(x c)	PROB. CONDICIONAL
P(c x)	PROB. POSTERIORI

P(D)	P(U)	P(R)
0,43	0,29	0,29



Naives Bayes

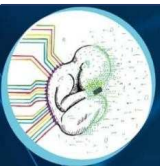
PROBABILIDADE CONDICIONAL

CALCULA A PROBABILIDADE DO TERMO PERTENCER À CLASSIFICAÇÃO.

P(c)	PROB. A PRIORE
P(x c)	PROB. CONDICIONAL
P(c x)	PROB. POSTERIORI

2ª - P(x C) probabilidades condicionais			C		
			Deisng	Usabilidade	Requisitos
X	x1	falta	0,018	0,020	0,021
	x2	tela	0,070	0,020	0,021
	x3	cores	0,035	0,020	0,021
	x4	identificar	0,018	0,020	0,021
	x5	menus	0,018	0,020	0,042

$$P(x/c) = \frac{nx+1}{fx(c)+fx(t)}$$



Naives Bayes

PROBABILIDADE CONDICIONAL

CALCULA A PROBABILIDADE DO TERMO PERTENCER À CLASSIFICAÇÃO.

$P(x c)$	PROB. CONDICIONAL
$P(c x)$	PROB. POSTERIORI

2º - $P(x C)$ probabilidades condicionais			C		
			Deisng	Usabilidade	Requisitos
X	x1	falta	0,018	0,020	0,021
	x2	tela	0,070	0,020	0,021
	x3	cores	0,035	0,020	0,021
	x4	identificar	0,018	0,020	0,021
	x5	menus	0,018	0,020	0,042

$$P(x/c) = \frac{nx+1}{fx(c)+fx(t)}$$

$$P(falta|Desing) = \frac{0*1+1}{19+38} = 0,018$$

$$P(tela|Desing) = \frac{3*1+1}{19+38} = 0,070$$

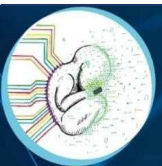
$$P(menus|requisitos) = \frac{1*1+1}{10+38} = 0,042$$



PROBABILIDADE À POSTERIORI

3º P(c X) Probabilidade a posteriori	log P(c)	log P(x1 c)	log P(x2 c)	log P(x3 c)	log P(x4 c)	log P(x5 c)	
P(Desing d8)	-0,37	-1,76	-1,15	-1,45	-1,76	-1,76	-8,24
P(Usabilidade d8)	-0,54	-1,69	-1,69	-1,69	-1,69	-1,69	-9,00
P(Requisitos d8)	-0,54	-1,68	-1,68	-1,68	-1,68	-1,38	-8,65

$P(c/x) = \log P(x|c) + \log P(x1|c) + \log P(x2|c) + \log P(x3|c) + \dots$



Naives Bayes

PROBABILIDADE À POSTERIORI

$P(x c)$	PROB. CONDICIONAL
$P(c x)$	PROB. POSTERIORI

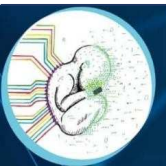
3ª $P(c X)$ Probabilidade a posteriori	$\log P(c)$	$\log P(x_1 c)$	$\log P(x_2 c)$	$\log P(x_3 c)$	$\log P(x_4 c)$	$\log P(x_5 c)$	
$P(\text{Desing} d8)$	-0,37	-1,76	-1,15	-1,45	-1,76	-1,76	-8,24
$P(\text{Usabilidade} d8)$	-0,54	-1,69	-1,69	-1,69	-1,69	-1,69	-9,00
$P(\text{Requisitos} d8)$	-0,54	-1,68	-1,68	-1,68	-1,68	-1,38	-8,65

$$P(c/x) = \log P(x|c) + \log P(x_1|c) + \log P(x_2|c) + \log P(x_3|c) + \dots$$

$$\log \frac{3}{7} = -0,37$$

$$\log 0,018 = -1,76 \quad \log 0,070 = -1,15 \quad \log 0,035 = -1,45$$

* Para evitar o underflow de ponto flutuante, o produto das probabilidades é substituído pela soma dos logaritmos das probabilidades



Naives Bayes

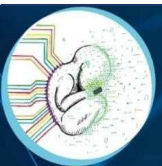


Algumas utilidades práticas:

- **Previsões em tempo real:** Por possuir uma velocidade relativamente alta e precisar apenas de poucos dados para realizar a classificação, o Naive Bayes pode ser utilizado para previsões em tempo real.
- **Classificação de textos/Filtragem de spam/Análise de sentimento:** Muito utilizado para filtragem de SPAM, Análise de Sentimento nas redes sociais(identificar se o usuário está feliz ou triste ao publicar determinado texto).



Matriz de confusão

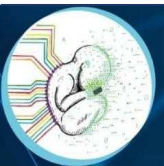


Matriz de Confusão

Valor Verdadeiro

Valor Previsto

		Valor Previsto	
		Positivo	Negativo
Valor Verdadeiro	Negativo	Verdadeiros Positivos	Falsos Negativos
	Positivo	Falsos Positivos	Verdadeiros Negativos



Matriz de Confusão

Em um conjunto de dados de 100 animais utilizados para prever se estes são ou não gatos.

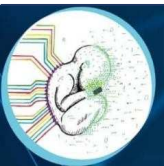
		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

		Classe esperada	
		Gato	Não é gato
Classe prevista	Gato	25 Verdadeiro Positivo	10 Falso Positivo
	Não é gato	25 Falso Negativo	40 Verdadeiro Negativo

Veridicamente 35 são gatos e 65 não são.



Validação do Modelo Métricas de Desempenho



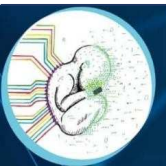
Matriz de Confusão

Acurácia

- É definida como sendo a fração de premissas corretas, representando a quantidade de acertos do classificador, verdadeiros positivos e verdadeiros negativos, em relação ao total calculado.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)



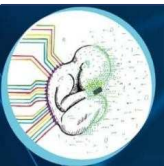
Matriz de Confusão

Precisão

Precisão ou valor preditivo positivo representa o quão precisas são as predições positivas de um modelo. Perfaz o número de casos verdadeiros positivos pelo total de instâncias positivas preditas.

$$Precision = \frac{TP}{TP+FP}$$

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)

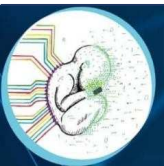


Métricas de Desempenho



A acurácia é a proximidade de um resultado com o seu valor de referência real. Dessa forma, quanto maior a acurácia, mais próximo da referência ou valor real é o resultado encontrado.

A precisão é o grau de variação resultante de um conjunto de medições realizadas. Dessa forma, quanto mais preciso um processo, menor é a variabilidade entre os valores encontrados.

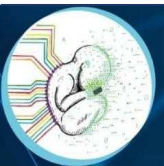


Recall

O Recall ou sensibilidade é considerado uma medida de completude da anterior. Verifica apenas qual a fração de positivos foram identificados pelo modelo, em relação ao total de instâncias positivas do conjunto original

$$Recall = \frac{TP}{FN + TP}$$

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)



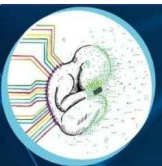
Matriz de Confusão

Média F1

A média F1 ou média harmônica combina os valores de precisão e sensibilidade em uma única fórmula.

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

		Valor Predito	
		Sim	Não
Real	Sim	Verdadeiro Positivo (TP)	Falso Negativo (FN)
	Não	Falso Positivo (FP)	Verdadeiro Negativo (TN)



Matriz de Confusão

Em um conjunto de dados de 100 animais utilizados para prever se estes são ou não gatos.

		Classe esperada	
		Gato	Não é gato
Classe prevista	Gato	25 Verdadeiro Positivo	10 Falso Positivo
	Não é gato	25 Falso Negativo	40 Verdadeiro Negativo

Veridicamente 35 são gatos e 65 não são.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

$$Recall = \frac{TP}{FN+TP}$$

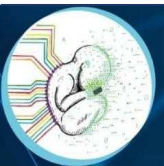
$$Precision = \frac{TP}{TP+FP}$$

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$



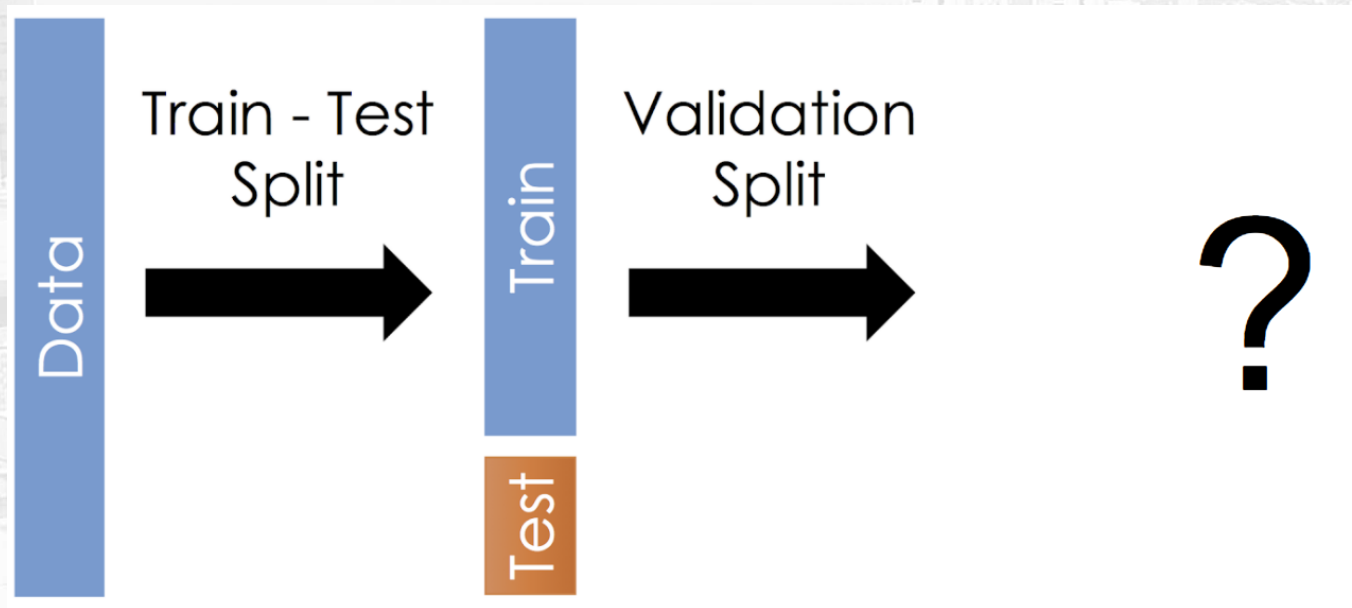
Validação dos dados

Dividindo dados de Treinamento e Teste



Dados de teste e treinamento

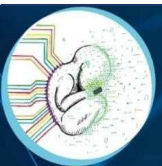
K folds





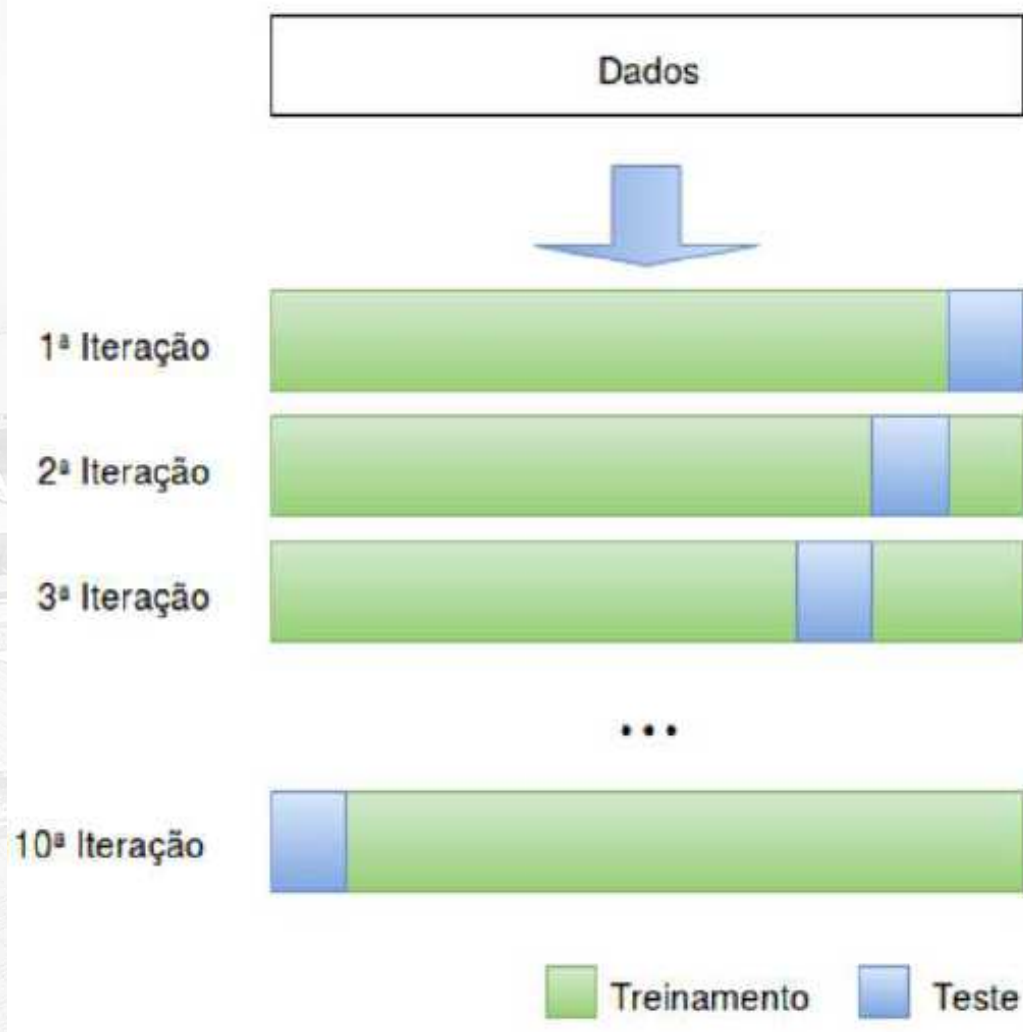
Validação dos dados

Cross Validation – Validação cruzada



Validação cruzada

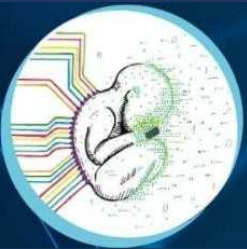
K folds





Trabalhando com o GOOGLE COLAB

<https://colab.research.google.com>



17ª SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA

Inteligência Artificial: A Nova Fronteira da Ciência Brasileira



Michel Ferreira Batista

Professor EBTT de Informática IFBA

michelfbatista@gmail.com

<https://sites.google.com/view/michelfbatista/>