Não é apenas futebol

Olá!

Me chamo Cássio Botaro

Estou aqui para falar sobre futebol.

Trabalho como desenvolvedor na **Nuveo.**

Você me encontra como @cassiobotaro

Aviso legal

- Não sou um cientista de dados
- Não é propaganda sobre a plataforma
- Propósito aqui é mostrar o case de forma didática

Todas as opiniões contidas aqui representam opinião do autor.

Nossa agenda

Contexto

Machine Learning 101

Análise de sentimentos

Nossa agenda

Sumarização

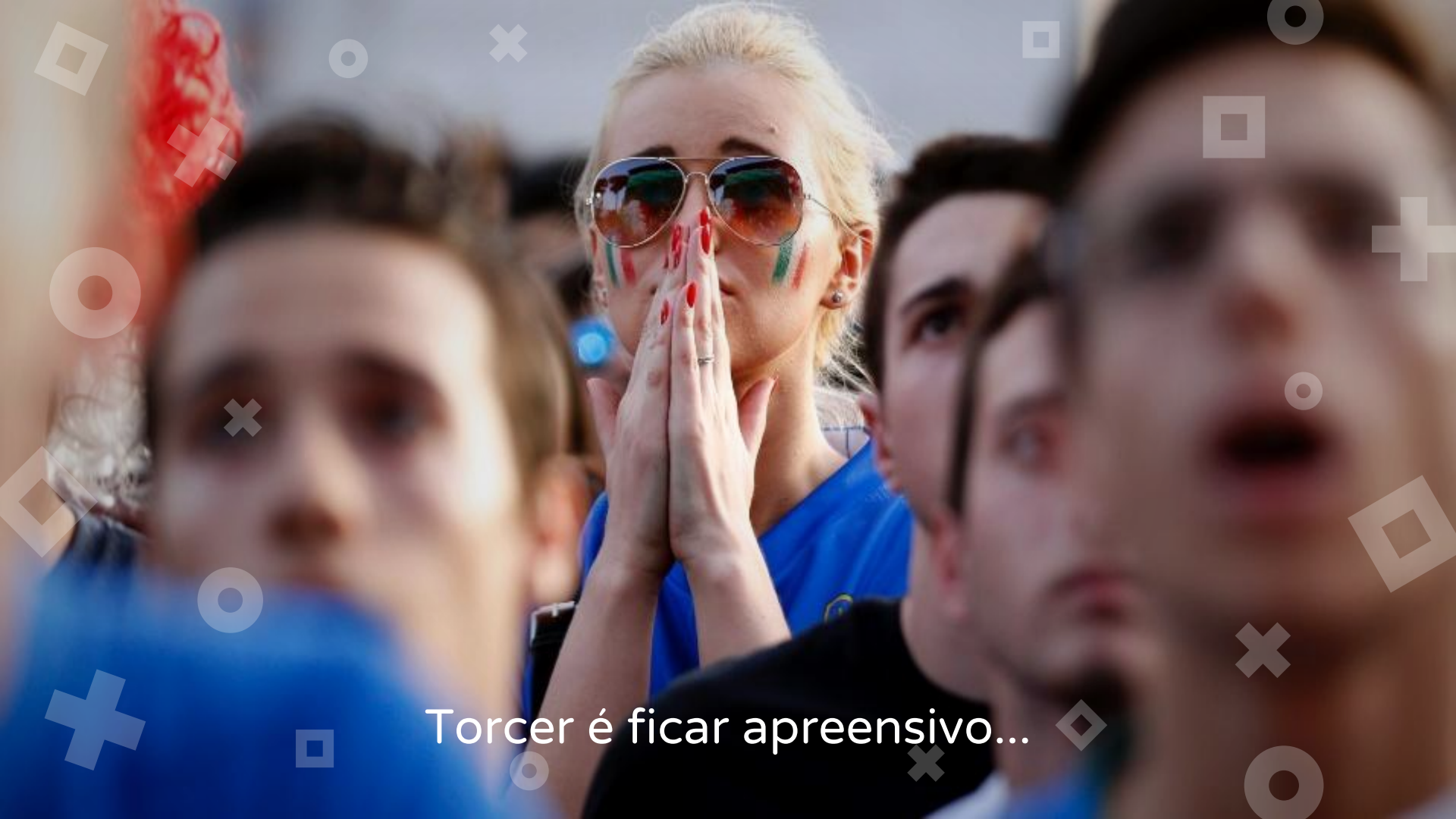
Planos futuros

Conclusão

Contexto

O que foi o sentibol?

Futebol não é apenas um jogo. É uma paixão.


Torcer é ficar apreensivo...

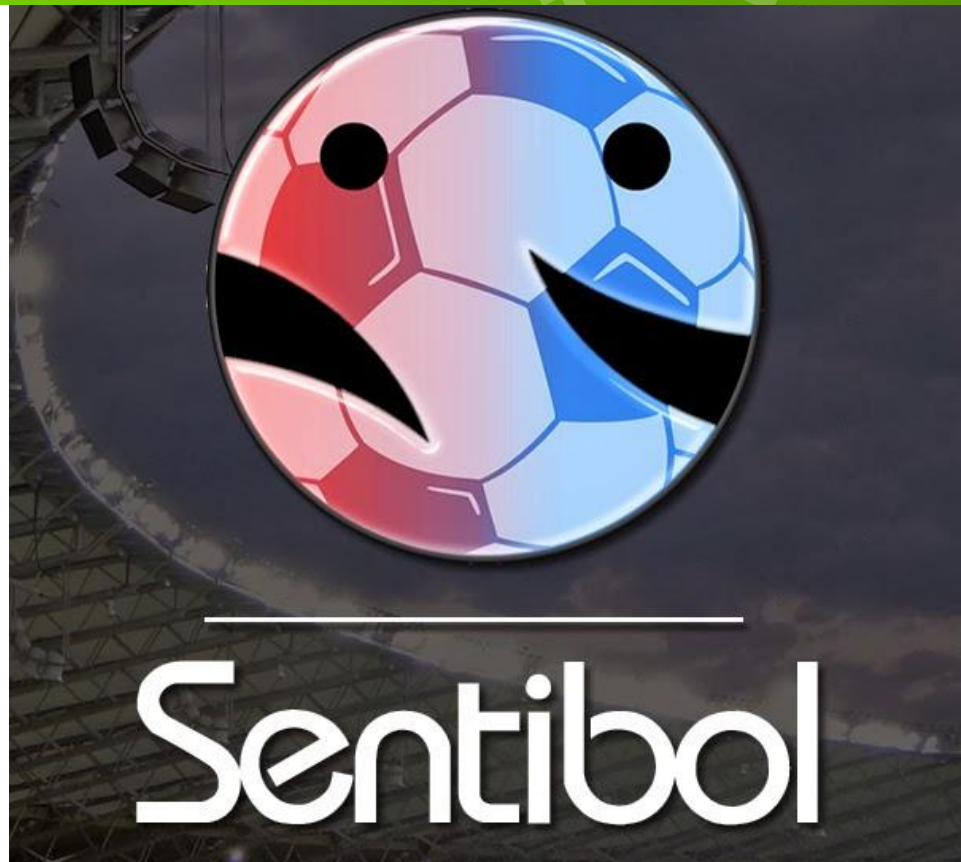

...e vibrar com a vitória!

Quando se trata de futebol...a coisa mais importante sobre futebol...é que não é apenas futebol.

Terry - escritor inglês

Sentibol

Uma plataforma para jornalistas, bloggers, entusiastas e afins com análises de dados referentes a plataforma twitter.



Sentibol - https://vimeo.com/88373037

Machine Learning 101

Conceitos necessários para o entendimento do resto da palestra.

Palavrinhas importantes

Labels

Informações de um dado, normalmente algo que você deseja saber.

Ex: spam ou não, sentimento positivo, negativo ou netro.

Features

Uma propriedade individual mensurável de um fenômeno observado.

Ex: peso, textura

Dataset

Conjunto de dados observados a fim de se extrair algum tipo de informação.

Classificação

Problema clássico do aprendizado de máquinas em que se busca a identificação de um subconjunto em um conjunto.

Ex: conjunto de maças e laranjas em um conjunto de frutas.

Acurácia

Medida de resultado de uma predição. Avaliação de uma classificação em termos práticos.

Predição

Prever um resultado com base em um conjunto de dados de testes.

Palavrinhas importantes

Classificador

Responsável por realizar a classificação segundo os critérios definidos em labels

Classificador por votação

Classificação utilizando múltiplos classificadores, tipicamente por votação de peso igualitário, mas podem apresentar pesos diferentes.

TF-IDF

É uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos.

N-gram

Sequência de n items de um texto ou fala.

Stopwords

Palavras que podem ser irrelevantes.

Análise de sentimentos

Determinar se um tweet é negativo, neutro ou positivo.

Tweets

o cruzeiro jogou muito bem

o cruzeiro nao jogou bem

cruzeiro jogou contra o sport

Remoção de stopwords

~~o~~ cruzeiro jogou ~~muito~~ bem

~~o~~ cruzeiro nao jogou bem

cruzeiro jogou contra ~~o~~ sport

Vetorização

	bem	contra	cruzeiro	jogou	nao	sport
cruzeiro jogou bem	0.673	0	0.523	0.523	0	0
cruzeiro não jogou bem	0.504	0	0.391	0.391	0.663	0
Cruzeiro jogou contra Sport	0	0.609	0.360	0.360	0	0.609

Treino

bem	contra	cruzeiro	jogou	nao	sport	Label
0.673	0	0.523	0.523	0	0	1
0.504	0	0.391	0.391	0.663	0	-1
0	0.609	0.360	0.360	0	0.609	0

Exemplo SVM

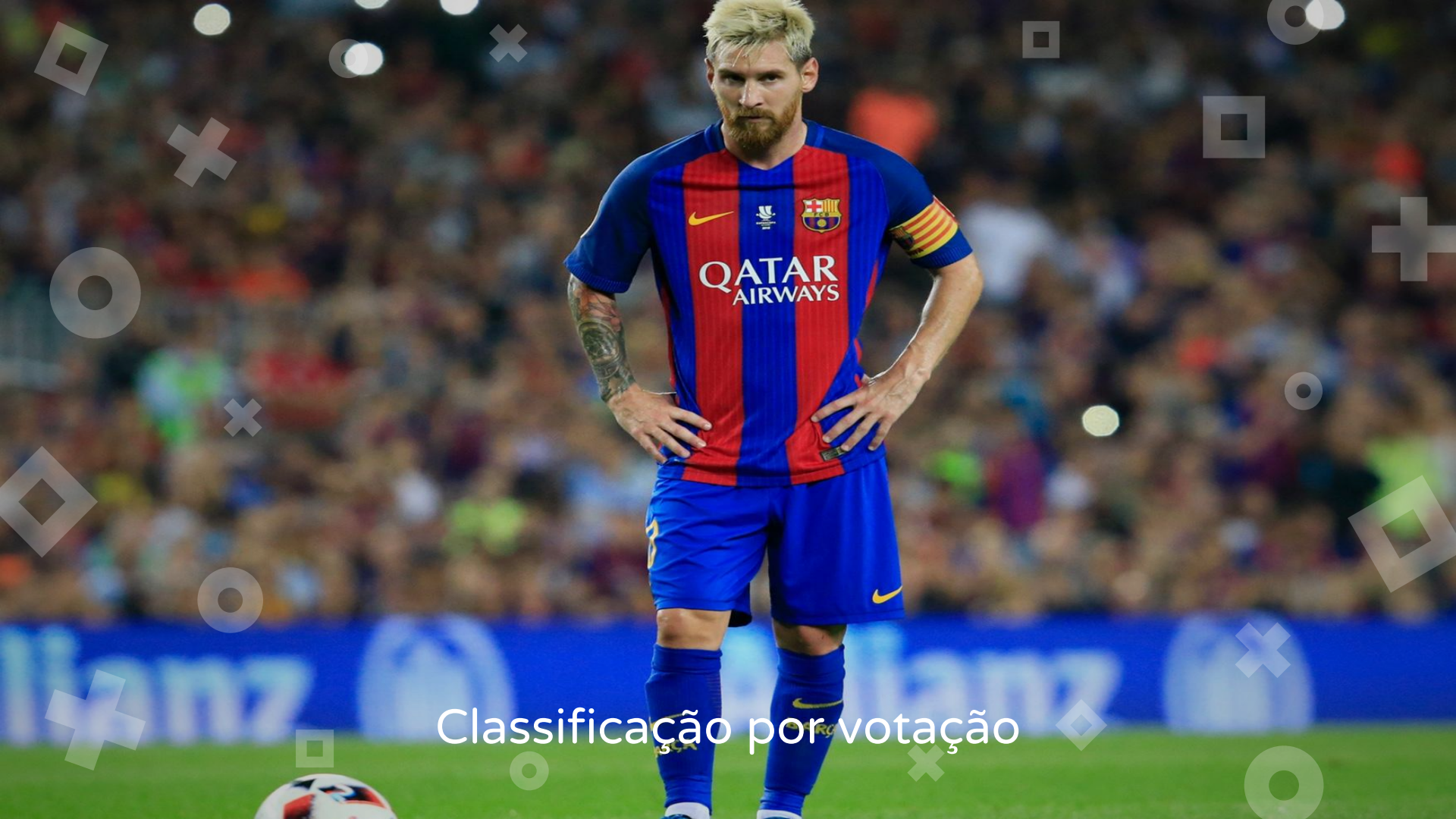
exemplo_svm.py

Exemplo Naive Bayes

exemplo_naive.py

Exemplo Sentiment Classification

exemplo_sentiment.py



Classificação por votação

SVM

Naive Bayes

Sentiment

classificaco_votacao.py

Sumarização

Determinar um elemento relevante em um conjunto

Exemplo Sumarização

exemplo_sumarizacao.py

Planos Futuros

O que vem por ai?

Planejamentos

- Associação entre notícias e análise
- Resolver o problema de uma base tendenciosa
- Resolver o problema de desambiguação de nomes
- Questionar os dados e descobrir mais informações

Conclusão

O que podemos concluir de tudo isso?

Estamos em um mar de informação e afogados na ignorância. - Carlos Filho

Obrigado!

Perguntas?

@cassiobotaro & cassiobotaro@gmail.com