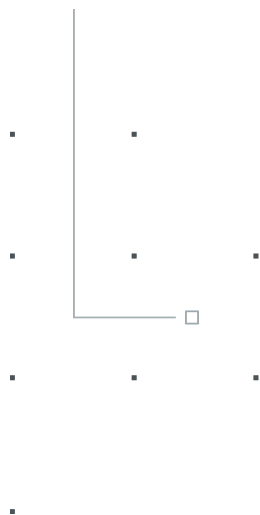


FIAP

NBA



# Classificação de Textos


Dheny R. Fernandes

- 1. Introdução**
- 2. O Problema de Classificação de Texto**
  1. Extração de características
  2. Pipeline de transformação
- 3. Análise de sentimentos**
- 4. Métricas de avaliação**
- 5. Demo**
- 6. Exercícios**

# Introdução



Isto é spam?



Estimado Cliente,  
Temos o prazer de informar que finalmente firmamos uma parceria com a Polícia Judiciária em resposta a ataques a sistemas bancários nos últimos anos.  
As medidas de segurança do seu cartão MULTIBANCO devem ser atualizadas o mais rápido possível para evitar novos abusos.

**ATUALIZAR AGORA**

Leva apenas 3 minutos,  
Obrigado por nos escolher!

MB WAY  
SIBS

Por favor não responda este email.  
MB WAY  
© 2021 SIBS Payments Solutions.

## Qual o gênero do autor?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

## Crítica de filme positiva ou negativa?



Incrivelmente desapontador



Cheio de personagens mirabolantes e uma sátira ricamente aplicada



O melhor filme de comédia já feito



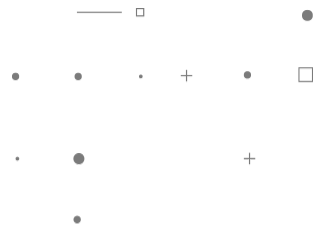
Foi patético. A pior parte foi a cena dentro do saguão.

Texto faz parte de nosso dia-a-dia. Assim, várias aplicações que, de alguma maneira, envolve texto e, principalmente, classificação de texto, surgem constantemente. Podemos elencar alguns exemplos:

- Atribuir categorias, tópicos ou gêneros a assuntos
- Classificação de mensagens textuais
- Identificação autoral
- Identificação de língua escrita
- Classificação de sentimento
- Chatbots.....

Vamos entender, então, o problema de classificação de texto





# O problema de classificação de texto



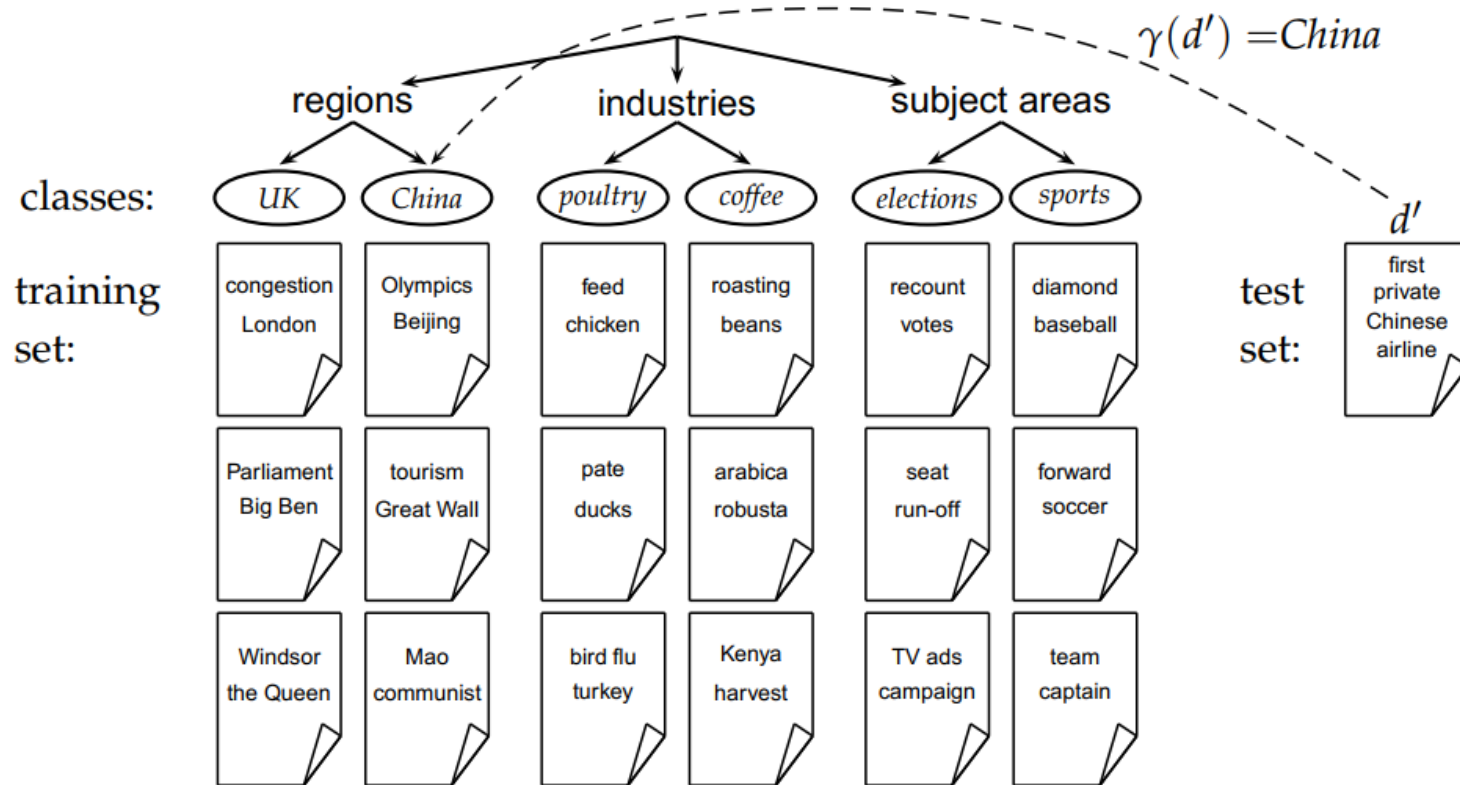
Seja:

- $d \in X$  um documento, em que  $X$  é o espaço de documentos
- $\mathbb{C} = \{c_1, c_2, \dots, c_j\}$  um conjunto fixo de Classes (categorias ou rótulos)
- $\{(d_1, c_1), (d_2, c_2) \dots, (d_m, c_m)\}$  um conjunto de treinamento de  $m$  documentos manualmente rotulados

Usando um [algoritmo de aprendizado](#), desejamos aprender uma função  $\gamma$  de classificação que mapeia documentos para classes:

- $\gamma = X \rightarrow \mathbb{C}$

Podemos representar o problema da seguinte maneira:



A definição dada estipula que um documento é **membro de uma, e apenas uma**, classe.

Porém, como deveria ser classificado um documento sobre as Olimpíadas de 2008?

Iremos ver uma abordagem para lidar com esse tipo de problema futuramente

Se  $d \in X$  é um documento que pertence ao espaço de documentos, como representa-lo de forma que um computador consiga interpretá-lo?

Como vimos na última aula, é preciso transformar esse documento em números.

O jeito como eu realizo essa transformação é denominado **modelo de representação**. Inicialmente, veremos dois tipos:

- Bag of Words
- TF-IDF

Vamos começar pelo BoW

O modelo Bag of Words usa um vetor de **contagens de palavras** para representar um documento.

$x = [1, 0, 1, 0, 4, 3, 10, 6, 7, \dots]$ , em que  $x_j$  é a contagem da palavra  $j$ .

O tamanho de  $x$  é determinado pelo **vocabulário**  $|\mathcal{V}|$ , que é o conjunto de todas as possíveis palavras no vocabulário

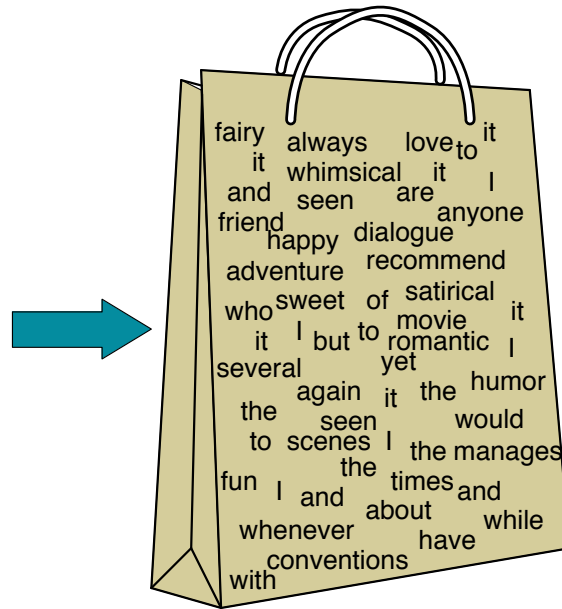
O modelo BoW é assim chamado por somente incluir informação sobre a **contagem de cada palavra**, e **não a ordem em que cada uma aparece**.

Ou seja, com o BoW, a semântica, contexto e as frases em si são ignorados.

Ainda assim, é **surpreendentemente efetivo** para classificação de texto

De modo ilustrado, é assim que funciona o BoW:

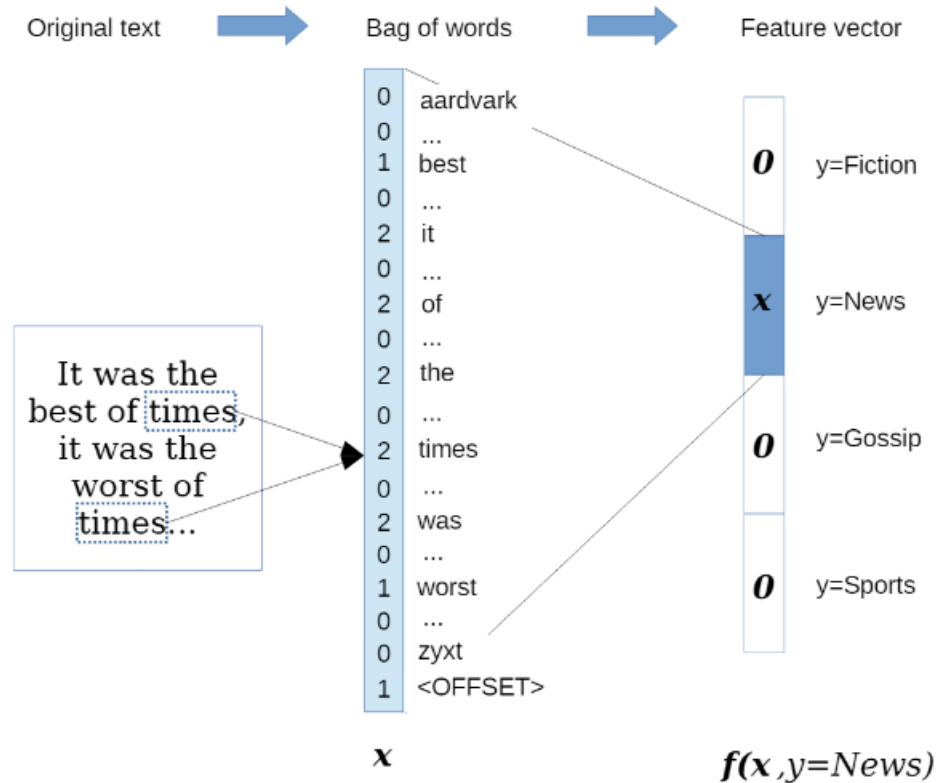
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



O mapeamento de características para output é ilustrado da seguinte forma:



Para entender o conceito de TF-IDF, responda: todas as palavras num documento são igualmente importantes?

Diante disso, introduzimos um conceito de frequência de termo que calcula a proporção de um termo num documento em relação ao número total de termos nesse documento.

Entretanto, um problema com pontuar frequências de palavras é que palavras muito frequentes começam a dominar no documento (pontuação alta), mas podem não conter muita “informação de conteúdo” para o modelo em relação a palavras raras que pertençam a domínios específicos.

Assim, introduzimos um mecanismo para **atenuar o efeito de termos que ocorrem muito nos dados** para tornar significativo a determinação de sua relevância.

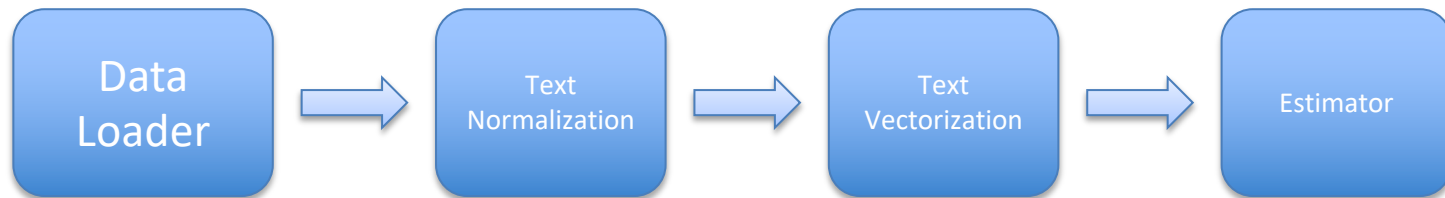
Chamamos esse mecanismo de **IDF** (*inverse document frequency*), que **mede a importância de um termo**.

Definimos, então, cada um dos termos:

- $tf = \frac{f_{t,d}}{\sum T,d}$
- $idf = \frac{N}{n_t}$
- $TFIDF = tf * idf$
- Em que:
  - $t$ : termo analisado
  - $d$ : documento analisado
  - $T$ : conjunto de termos presente no documento  $d$
  - $N$ : número total de documentos
  - $n_t$ : número de documentos que contém o termo  $t$

Até agora, vimos uma série de transformações que podemos fazer nos dados textuais, mas como organizar isso de modo a construir um pipeline a fim de realizar essas transformações?

Um simples pipeline pode consistir das seguintes etapas:



O processo de normalização de texto pode ser composto pelas seguintes partes:

- Tokenização
- Remoção de stop-words
- Remoção de acentuação e pontuação
- Lematização e/ou stemização
- lowercasing

No código, eu apresento o processo de remoção de acentuação e pontuação.

# Análise de Sentimentos

A análise de sentimentos nada mais é que uma forma de classificação de texto. Vimos esses exemplos anteriormente sobre avaliação de filmes. Nesse caso, o objetivo da classificação de texto consiste em, dado um novo review, predizer seu sentimento.



Incrivelmente desapontador



Cheio de personagens mirabolantes e uma sátira ricamente aplicada



O melhor filme de comédia já feito

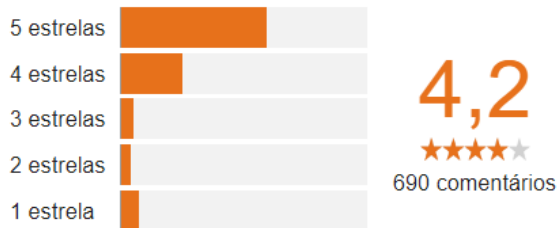


Foi patético. A pior parte foi a cena dentro do saguão.



Aqui temos mais um exemplo. Qual é o problema que temos aqui?

## Resenhas

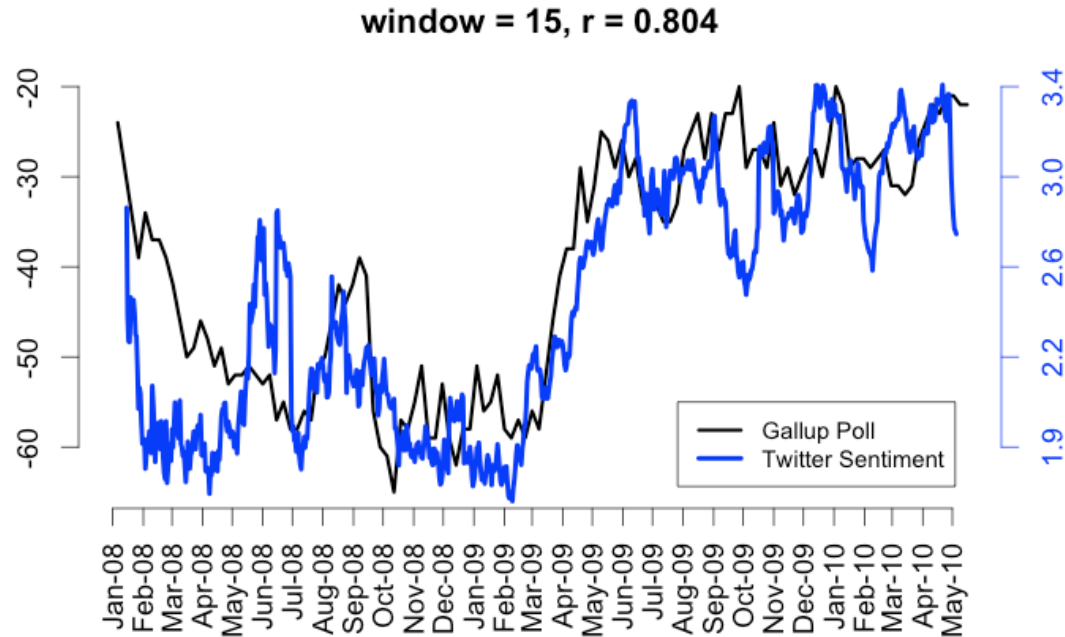


★★★★★ **ATENÇÃO NÃO IMPRIME PAPEL FOTOGRAFICO** , – 22 de Junho de 2018

sergioluis – Resenha fornecida por [E](#) [epson.com.br](#)  
22 de Junho de 2018

ATENÇÃO NÃO IMPRIME PAPEL FOTOGRAFICO, APESAR DA IMAGEM DA CAIXA E FOLDER PARECER UMA FOTO, O VENDEDOR DA CASAS BAHIA COITADO INFORMOU ERRADO, ATE TROQUEI A IMPRESSORA PENSANDO QUE ESTAVA COM DEFEITO E FUI NA ASSIST TECNICA A QUAL ME ORIENTOU EM TROCAR POIS ESTAVA NA GARANTIA, CONCLUSÃO FIQUEI COM PREJUIZO. DEFEITO FICA COM VARIOS BARRAMENTOS, HORRIVEL. TAMBEM NÃO TEM SUPORTE PARA O PAPEL NA SAIDA, PENSO QUE O CLIENTE DEVERIA SER INFORMADO QUANDO COMPREI MOSTREI UMA FOTO A QUAL PRECISAVA E FUI ORIENTADO QUE ESTA IMPRESSORA SERVERIA, E QUE TAMBEM DIMINUIRIA O TOTAL DE IMPRESSÃO EM APENAS 1000 FLS, A QUAL NEM IMPRIMI 15 FLS EM TESTES JÁ ESTA NO FINAL O NIVEL DE TINTA??

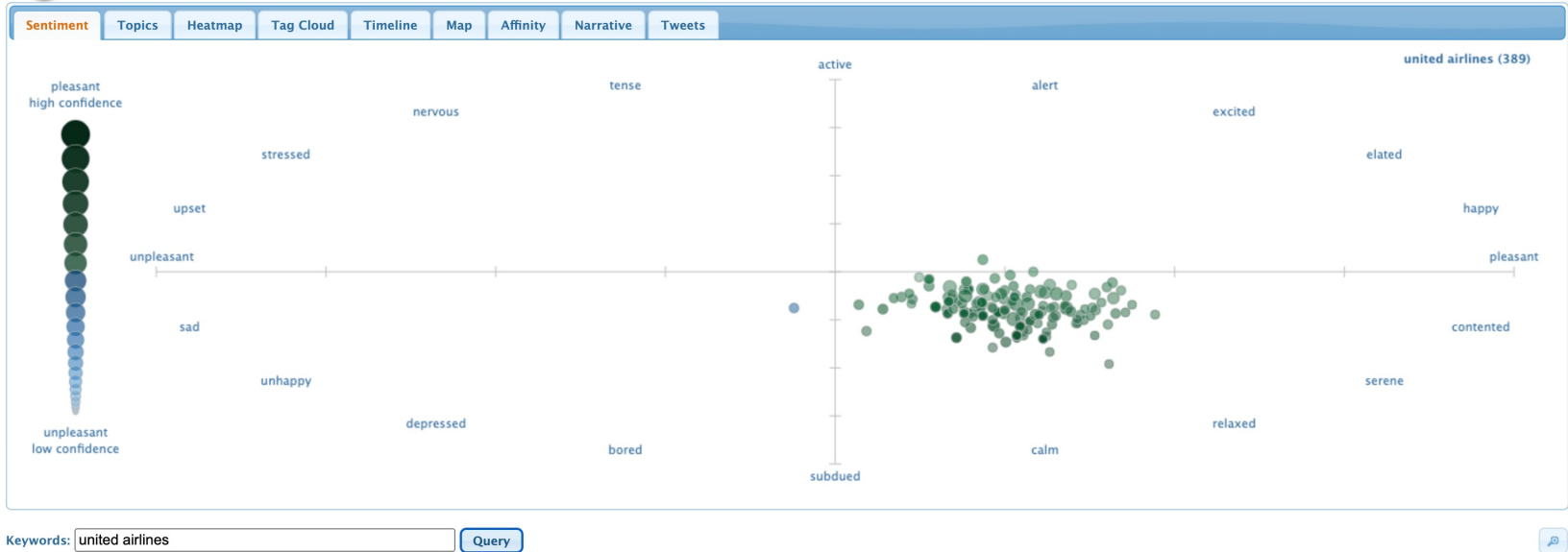
Esse gráfico compara o sentimento extraído de tweets com a pesquisa de confiança do consumidor realizada pela Gallup Poll:



## Curiosidade: [Twitter Sentiment App](#)



sentiment viz  
Tweet Sentiment Visualization

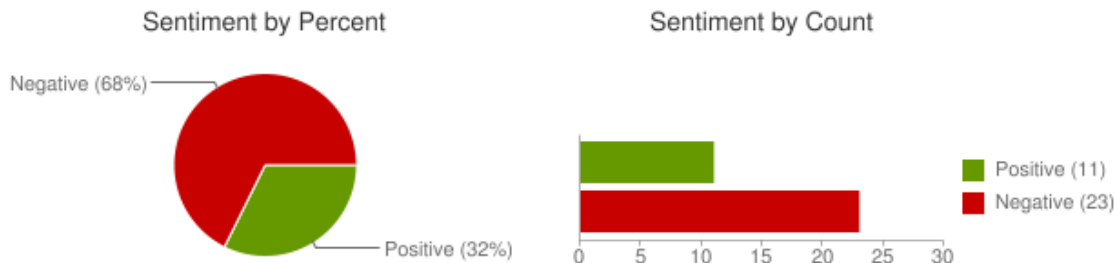


## Mas já foi assim o sentimento da United Airlines:

Type in a word and we'll highlight the good and the bad

[Save this search](#)

### Sentiment analysis for "united airlines"



jljacobson: OMG... Could **@United airlines** have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.  
Posted 2 hours ago

12345clumsy6789: I hate **United Airlines** Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?  
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 **united airlines** and 24seven to an exotic destination. <http://t.co/Z9QloAjF>  
Posted 2 hours ago

CountAdam: FANTASTIC customer service from **United Airlines** at XNA today. Is tweet more, but cell phones off now!  
Posted 4 hours ago

E por qual razão Análise de Sentimentos é importante?

Bem, ela faz parte de nosso dia-a-dia:

- **Filmes**: o que a crítica e o público tem dito acerca do novo lançamento?
- **Produtos**: o que as pessoas pensam sobre o novo Iphone?
- **Sentimento público**: como está a confiança do consumidor?
- **Política**: o que as pessoas pensam a respeito de um candidato?
- **Predição**: qual a tendencia de mercado?

# Métricas de Avaliação



You can't  
manage what  
you can't  
measure

-Peter Drucker

Precisamos de maneiras de mensurar a qualidade de predição de nosso algoritmo. Em classificação, as métricas mais comuns, e as mais simples, são **Acurácia**, **Precision**, **Recall** e **F1 Score**. Elas são obtidas a partir da [Matriz de Confusão](#), apresentada abaixo:

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN



A acurácia é dada pela seguinte função:

$$acc = \frac{tp + tn}{(tp + tn + fp + fn)}$$

Precision (acurácia das predições positivas) pode ser calculado da seguinte maneira:

$$P = \frac{tp}{(tp + fp)}$$

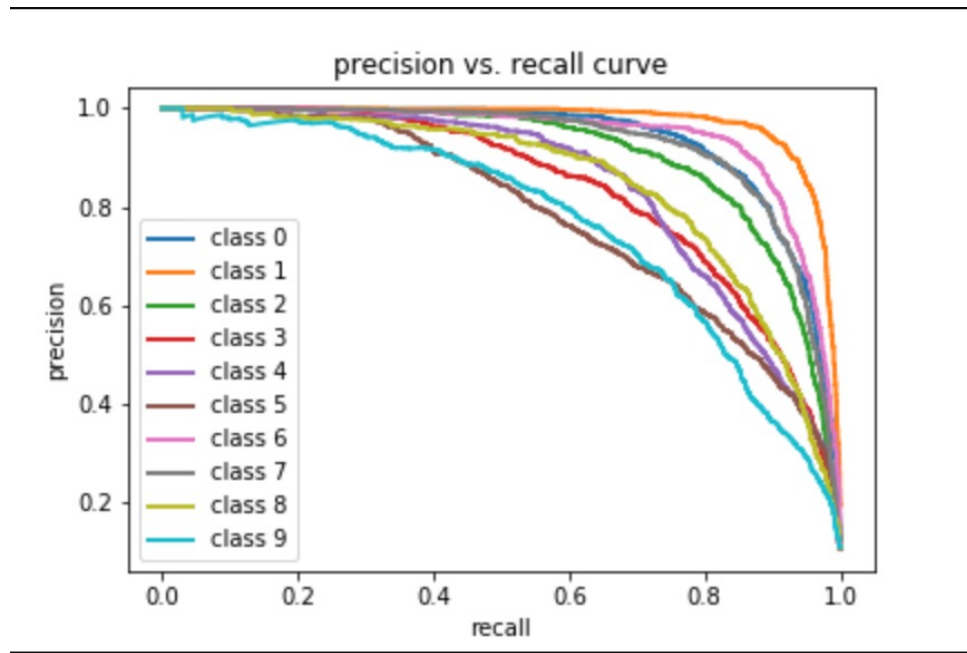
Já o Recall (taxa de amostras positivas corretamente encontradas pelo classificador) é calculado da seguinte forma:

$$R = \frac{tp}{(tp + fn)}$$

**Precision** e **Recall** tendem a ser inversamente proporcionais, como apresentado no gráfico ao lado.

Para minimizar esse problema, existe a métrica **F1-Score**, que leva em consideração tanto Precision quanto Recall.

$$F1 = \frac{2PR}{P + R}$$




# Demo e Exercícios



# Obrigado!

profdheny.fernandes@fiap.com.br

 /dhenyfernandes

FIAP MBA<sup>+</sup>

Copyright © 2022 | Professor Dheny R. Fernandes

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.

FIAP