

MINERAÇÃO DE OPINIÕES DESAFIOS E TÉCNICAS

Francisco Assis Ricarte Neto
Rafael Tôrres Anchiêta

Agenda

- Motivação e Contexto
- Mineração de Opiniões
 - Nomenclatura e Definições
 - Áreas Correlatas
 - Níveis de Granularidade
 - Etapas da AS
 - Desafios
- Técnicas
 - Introdução Inteligência Artificial
 - Aprendizagem de Máquina
 - Processamento de Linguagem Natural
 - Stemmers, Pos-Taggers, Parsers
- Prática

MOTIVAÇÃO E CONTEXTO

Imaginem

- Há 40, 30, ou até 20 anos atrás como fazíamos para saber a procedência de:
 - Hotéis
 - Carros
 - Passeios Turísticos

Imaginem

- Há 40, 30, ou até 20 anos atrás como fazíamos para saber a procedência de:
 - Hoté
 - Carrc
 - Pass



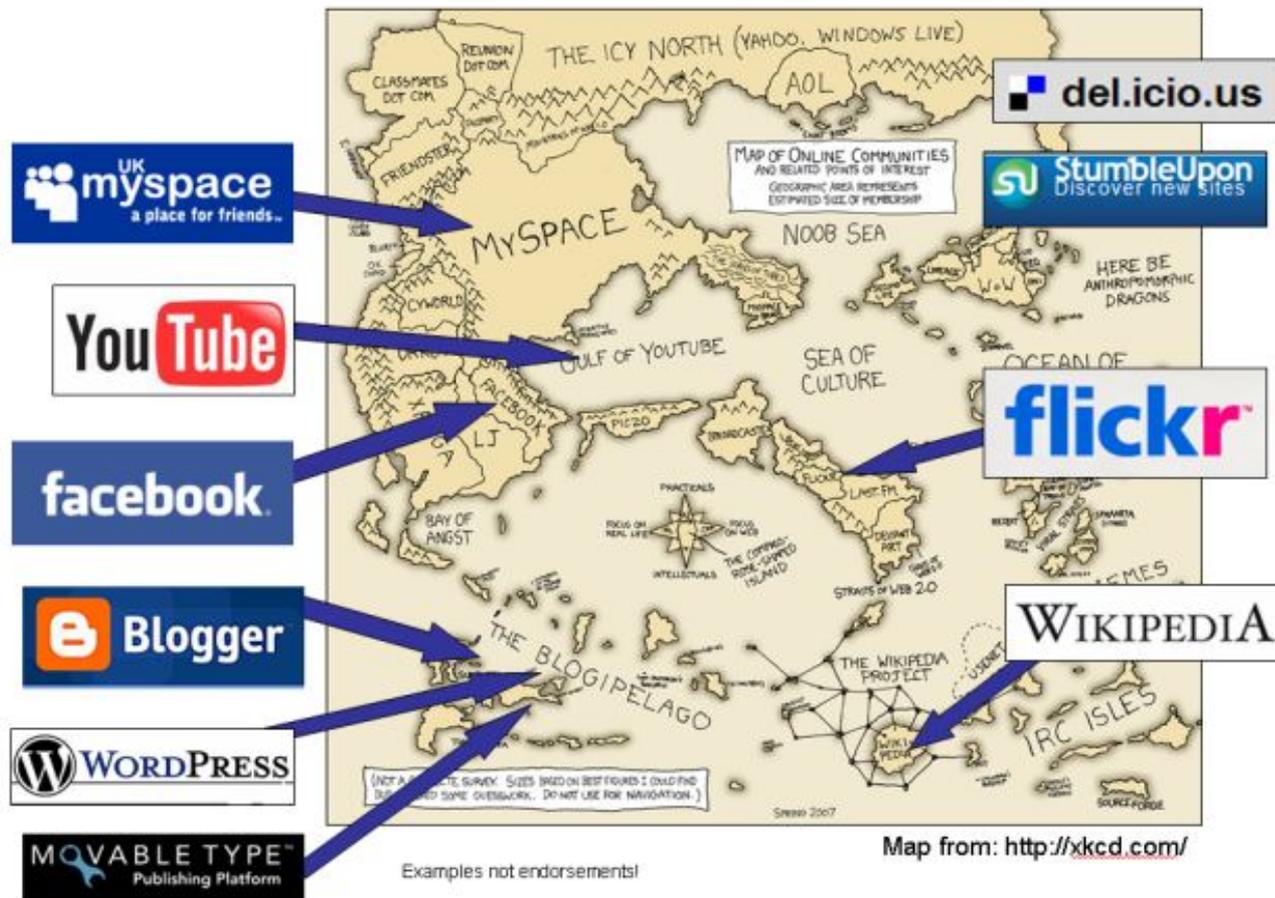
Motivação

- O que temos?
 - Milhares de Opiniões disponíveis na Web sobre os mais diversos assuntos
 - Ex.: produtos, viagens, hotéis, política, ideologias, serviços em geral.

Motivação

Motivação

Internet Communities: post 'Web 2.0'



Motivação

Customer Reviews

Panasonic KX-TG1032S DECT 6.0 Expandable Digital Cordless Phone System with 2 Handsets

167 Reviews

Average Customer Review: ★★★★★ (167 customer reviews)

Search Customer Reviews: Only search this product's reviews

5 star: (94)
4 star: (44)
3 star: (11)
2 star: (9)
1 star: (9)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

114 of 121 people found the following review helpful:

★★★★★ **Best DECT6.0; renders 5.8GHz obsolete!**

We'd looked long and hard at the latest 5.8GHz Unidens (TRU9300 and TRU9400) and Panasonic (6000 series) and hadn't been exactly overwhelmed by the range and voice quality. Battery life was a joke.

The we heard about the new Panasonic DECT 6.0 phones. In a word: WOW! Dramatically better range, sound quality, and rated talk time than any of the 5.8ghz...

[Read the full review >](#)

Published 12 months ago by silphium

> See more [5 star](#), [4 star](#) reviews

The most helpful critical review

33 of 38 people found the following review helpful:

★☆☆☆☆ **Bad QC, technical support even worse**

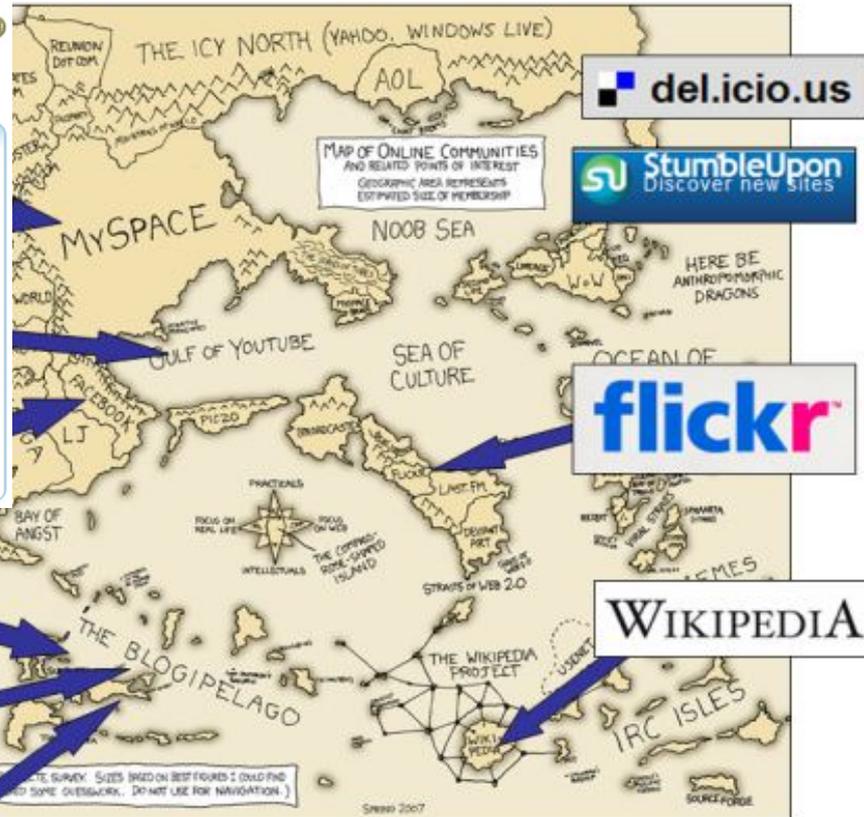
Less than a month after the purchase, the base-unit button for retrieving the message stops working. Manufacturer's technical/customer support is practically non-existent. The limited 1 yr warranty provides only a replacement with a refurbished (NOT NEW!) product, and only if the customer mails the phone to their TX center, at own expense. Panasonic, shame on you!!!

Published 7 months ago by Damir Vidovic

> See more [3 star](#), [2 star](#), [1 star](#) reviews

Vs.

Internet Communities: post 'Web 2.0'



Map from: <http://xkcd.com/>

Examples not endorsements!

Motivação

Customer Reviews

Panasonic KX-TG1032S Dect 6.0 Expandable Digital Cordless Phone System with 2 Handsets

167 Reviews

Average Customer Review **★★★★☆** (167 customer reviews)

Search Customer Reviews

Share your thoughts with other customers

Only search this product's reviews

Create your own review

5 star: (94)
4 star: (44)
3 star: (11)
2 star: (9)
1 star: (9)

The most helpful favorable review

114 of 121 people found the following review helpful:

★★★★★ **Best DECT6.0; renders 5.8GHz obsolete!**

We'd looked long and hard at the latest 5.8GHz Unidens (TRU9300 and TRU9400)

By Cory #2
Kansas City, MO
on January 28, 2011

By chkcutelets
on January 28, 2011

The most helpful critical review

33 of 38 people found the following review helpful:

★☆☆☆☆ **Bad QC, technical support even worse**

Less than a month after the purchase, the base-unit button for retrieving the

★★★☆☆

“ Pretty good recipe, although I did make one substitution. Instead of dark chocolate I used beef jerky. ”

5 people found this review Helpful.
Was this review helpful to you? [Yes](#) | [No](#)

★★★★★

“ This is a great recipe. A great twist is adding 3 scoops of ice cream, whipped cream, nuts, sprinkles, and hot fudge. Before digging in, take the one oz dark chocolate and place it i the trash. ”

6 people found this review Helpful.
Was this review helpful to you? [Yes](#) | [No](#)

Internet Communities: post 'Web 2.0'



Map from: <http://xkcd.com/>

Motivação

Customer Reviews

Panasonic KX-TG1032S Dect 6.0 Expandable Digital Cordless Phone System with 2 Handsets

167 Reviews

Average Customer Review **★★★★☆** (167 customer reviews)

Search Customer Reviews

Share your thoughts with other customers

Only search this product's reviews

Create your own review

The most helpful favorable review

114 of 121 people found the following review helpful:

★★★★★ Best DECT6.0; renders 5.8GHz obsolete!
We'd looked long and hard at the latest 5.8GHz Unidens (TRU9300 and TRU9400)

 By Cory #2
Kansas City, MO
on January 28, 2011

The most helpful critical review

33 of 38 people found the following review helpful:

★☆☆☆☆ Bad QC, technical support even worse
Less than a month after the purchase, the base-unit button for retrieving the

★ ★ ★ ☆ ☆

“ Pretty good recipe, although I did make one substitution. Instead of dark chocolate I used beef jerky. ”

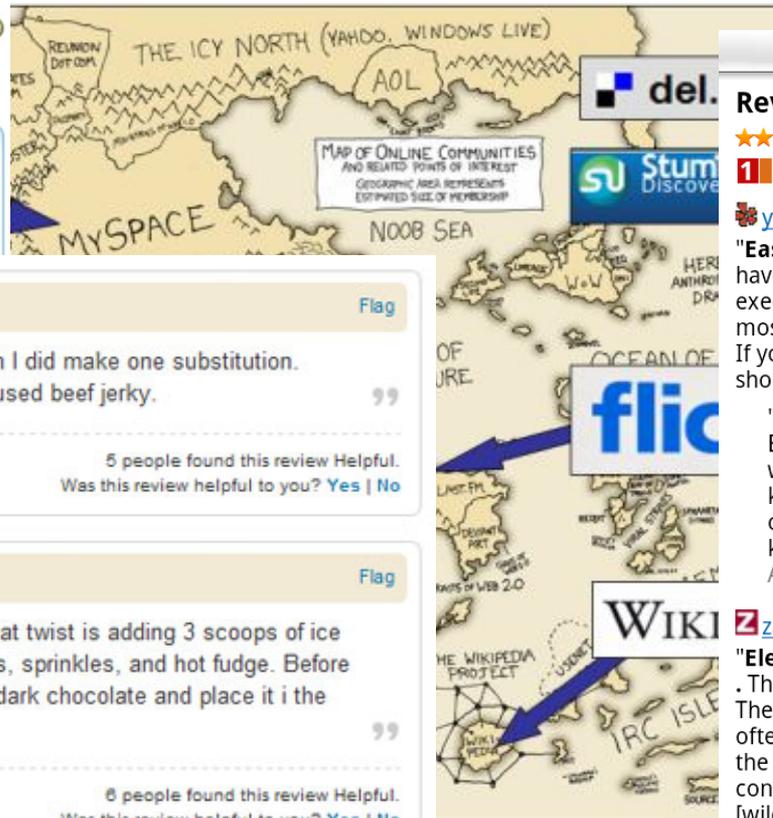
5 people found this review Helpful.
Was this review helpful to you? [Yes](#) | [No](#)

★ ★ ★ ★ ★

“ This is a great recipe. A great twist is adding 3 scoops of ice cream, whipped cream, nuts, sprinkles, and hot fudge. Before digging in, take the one oz dark chocolate and place it i the trash. ”

8 people found this review Helpful.
Was this review helpful to you? [Yes](#) | [No](#)

Internet Communities: post 'Web 2.0'



Map from: <http://xkcd>



Reviews from around the web

★★★★★ 173 reviews

1 3 4 5 stars

 [yelp.com - 302 reviews](#) ★★★★★

"Easily the BEST fine dining experience I have ever had. Amazing ingredients, amazing execution, amazing service, amazing wine...and most surprisingly...amazing value for fine dining. If you haven't been to **Acquerello**, you really should try it. ..." - Apr 18, 2010 - [Full review](#)

"Tobacco flavored cheese?? I'm there !! Excellent food and service. The sommelier was very young, but surprisingly knowledgeable and he spent a good amount of time explaining the wine pairings. The kitchen was very accommodating with ..." - Apr 15, 2010 - [Full review](#)

 [zagal.com - 45 reviews](#) ★★★★★

"Elegant service, elegant room and elegant .." The sommelier really knows his Italian whites. There was a certain timidity in the food which often happens with Italian haute cuisine. Had the risotto with strawberries - very good, considering the real dish is made with fragolini [wild ...]" xtinem - Sep 28, 2010 - [Full review](#)

 [reviews.opentable.com - 27 reviews](#) ★★★★★

"The food is average when you compare it to



Examples not endorsements!

Motivação

Customer Reviews

Panasonic KX-TG1032S Dect 6.0 Expandable Digital Cordless Phone System with 2 Handsets

167 Reviews

Average Customer Review: ★★★★★ (167 customer reviews)

Search Customer Reviews: Only search this product's reviews

5 star: (94)
4 star: (44)
3 star: (11)
2 star: (9)
1 star: (9)

Share your thoughts with other customers

Create your own review

The most helpful favorable review

114 of 121 people found the following review helpful:

★★★★★ **Best DECT6.0; renders 5.8GHz obsolete!**

We'd looked long and hard at the latest 5.8GHz Unidens (TRU9300 and TRU9400)

The most helpful critical review

33 of 38 people found the following review helpful:

★☆☆☆☆ **Bad QC, technical support even worse**

Less than a month after the purchase, the base-unit button for retrieving the

By Cory #2
Kansas City, MO
on January 28, 2011



66 Pretty good recipe, although I did make a mistake. Instead of dark chocolate I used milk chocolate.

Excelente

19 de Outubro de 2012

Custo/benefício: Excelente

Qualidade: Excelente

Ótimo livro

Avallado por: hugotik

Ótimo livro, muito bom!! História envolvente e bacana.

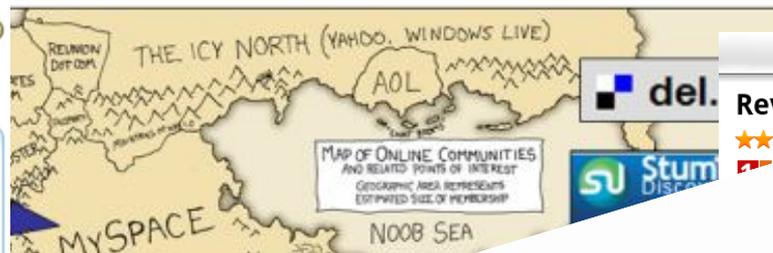
Essa avaliação foi útil?

Útil 0 Inútil 0

8 people found this review Helpful. Was this review helpful to you? Yes | No

Examples not endorsements!

Internet Communities: post 'Web 2.0'



12:35 AM

Reviews from around the web

★★★★★ 170

★★★★★

Experience I had at this restaurant was amazing. The wine pairings were perfect. The service was excellent. The food was delicious. I highly recommend this restaurant for fine dining. You really can't go wrong here!!

Postar comentário

...a good amount of wine pairings. The service was very accommodating with ..."

15, 2010 - [Full review](#)

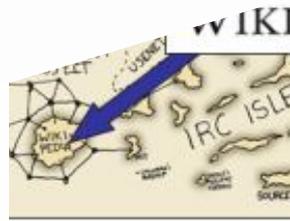
zagat.com - 45 reviews ★★★★★

"Elegant service, elegant room and elegant food. The sommelier really knows his Italian whites. There was a certain timidity in the food which often happens with Italian haute cuisine. Had the risotto with strawberries - very good, considering the real dish is made with fragolini [wild ...]"

28, 2010 - [Full review](#)

reviews.opentable.com - 27 reviews ★★★★★

"The food is average when you compare it to other restaurants in the area. The service is excellent. The wine pairings were perfect. The sommelier really knows his Italian whites. There was a certain timidity in the food which often happens with Italian haute cuisine. Had the risotto with strawberries - very good, considering the real dish is made with fragolini [wild ...]"



Map from: <http://xkcd.com>

Motivação

Customer Reviews

Panasonic KX-TG1032S Dect 6.0 Expandable Digital Cordless Phone System with 2 Handsets

167 Reviews

Average Customer Review **★★★★☆** (167 customer reviews)

Search Customer Reviews

Share your customer review with other

Only search this product's reviews

5 star: (94)

4 star: (44)

3 star: (11)

2 star: (9)

1 star: (9)

Share your customer review with other

Only search this product's reviews

Create

The most helpful favorable

114 of 121 people found the following helpful

★★★★★ Best DECT6.0; r 5.8GHz obsolete!

We'd looked long and hard for 5.8GHz Unidens (TRU930



By Cor
Kansa
on Ja

Já chega. ★★★★★
de Matheus S. Andrade - Versão 10.8.2 - 29-Sep-2012

Estava segurando, mas não dá mais. Não aguento mais os problemas com o Wi-Fi, não aguento mais a bateria acabar tão rápido, não aguento mais tanto travamento, não aguento mais aquele maldito círculo colorido... Quando comprei um Mac, o que eu sabia era que o Mac OS X raramente travava e tinha pouquíssimos problemas. O Snow Leopard realmente era assim, mas o Mountain Lion... Cada atualização é uma esperança de que os problemas sumam, mas até agora nada! Eu só não faço downgrade pro Snow Leopard porque, além de ter pago 40 reais nessa coisa, ele realmente tem coisas interessantes que não quero abrir mão, como o iCloud, os aplicativos em tela cheia e a Central de Notificações. Acho que a Apple está começando a sentir a falta de Steve Jobs, vide o problema com o Maps. Infelizmente...

158 de 176 cliente(s) acharam esta resenha útil
Esta resenha foi útil? [Sim](#) | [Não](#)



de erlon79 - Versão 10.8 - 27-Jul-2012

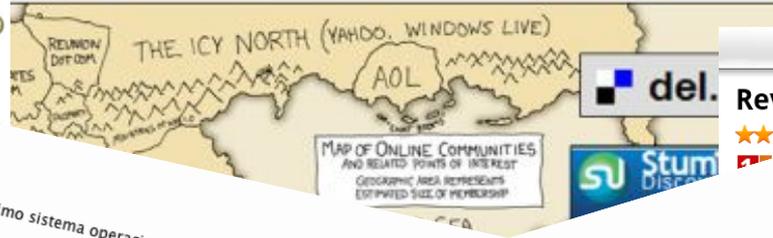
Me arrependi de ter comprado imediatamente. Está com problemas! ★★★★★

Eu era tão feliz com o Snow leopard e não sabia... com o Lion veio incompatibilidade com vários aplicativos importantes... e finalmente o Mountain Lion acabou também com minha navegação tanto com o Safari quanto com o Mozilla. Impressionantemente o Facebook faz o login mas depois não carrega mais nenhuma página e nem recarrega a página atual. O Twitter nem o login faz. Fica eternamente em andamento. Já tentei configurar opções de segurança e privacidade e nada... Já formatei e reinstalei o Mountain Lion também. Não compre agora, se não quiserem perder a navegação em alguns sites. Vou me comunicar com a Apple pra ver se eu resolvo e depois informo o resultado.

97 de 121 cliente(s) acharam esta resenha útil
Esta resenha foi útil? [Sim](#) | [Não](#)

Qualidade
Excelente

Internet Communities: post 'Web 2.0'



Reviews from around the web

★★★★★ 170

★★★

Experience I
... amazing
... wine...and
... for fine dining.
... you really
... [I review](#)

... there!!
... mmelier

... a good amount
... ine pairings. The
... modating with ..." -
... [eW](#)

★★★★★

... nt room and elegant ..
... knows his Italian whites.
... idity in the food which
... lian haute cuisine. Had
... berries - very good,
... ish is made with fragolini
... 28, 2010 - [Full review](#)

[ie.com](#) - 27 reviews ★★★★★

... when you compare it to

Todas as resenhas deste usuário >

Relatar uma preocupação >

Todas as resenhas deste usuário >

Relatar uma preocupação >

Todas as resenhas deste usuário >

Relatar uma preocupação >

Examples not endorsed

Motivação

- E pra que servem as Opiniões?
 - Opiniões são geralmente expressões subjetivas que descrevem sentimentos, avaliações de pessoas a cerca de entidades, eventos e suas propriedades.” [Liu, 2010]
 - Auxilia no processo de tomada de decisão
 - Ajuda empresas/firmas em processos de marketing

Por que monitorar?

- Obter o feedback dos clientes sobre a marca
- Poder pensar em novas estratégias
- Reverter uma opinião negativa sobre a empresa
- Conhecer seu público-alvo

Motivação

- Segundo o Ibope Media, somos 105 milhões de internautas tupiniquins (out-2013), sendo o Brasil o 5º país mais conectado [\[tobequarany, 2014\]](#)
- 57,2 milhões de usuários acessam regularmente a Internet
 - 38% das pessoas acessam à web diariamente;
 - 10% de quatro a seis vezes por semana;
 - 21% de duas a três vezes por semana;
 - 18% uma vez por semana.
- Assim, 87% dos internautas brasileiros entram na internet pelo menos uma vez por semana

Motivação

- Comercio Eletrônico

- Em 2008 foram gastos R\$ 8,2 bilhões em compras on-line.
- Em 2009, mesmo com crise, foram gastos R\$ 10,6 bilhões.
- 2010 fechou com R\$ 14,8 bilhões, atingindo 1/3 de todas as vendas de varejo feitas no Brasil.
- O último dado é de 2012, quando foram gastos 22,5 bilhões. Ainda assim, apenas 20% dos internautas brasileiros fazem compras na internet; aqueles que ainda não compram, não o fazem por não considerar a operação segura (69%) ou porque não confiam na qualidade do produto (26%).

Motivação

- Publicidade on-line
 - A internet se tornou o terceiro veículo de maior alcance no Brasil, atrás apenas de rádio e TV. 87% dos internautas utilizam a rede para pesquisar produtos e serviços. Antes de comprar, 90% dos consumidores ouvem sugestões de pessoas conhecidas, enquanto 70% confiam em opiniões expressas online.

Contexto

- Entretanto, encontrar opiniões na web e minerá-las é uma tarefa bastante exaustiva
 - Grande quantidade de *reviews*
 - Fontes relevantes
 - Opiniões “escondidas” em textos bastante extensos
 - Extrair sentenças opinativas e interpretá-las



MINERAÇÃO DE OPINIÕES

Mineração de Opiniões

- “Mineração da Opinião ou Análise de Sentimento (AS) é o estudo computacional de opiniões, sentimentos e emoções expressas acerca de entidades, eventos e seus atributos, que estão em um texto.” [Liu, 2010]
- “Mineração da Opinião é o problema de identificar opiniões expressadas sobre um determinado assunto e avaliar a polaridade dessa opinião.” [Tsytsarau and Palpanas, 2011]

Mineração de Opiniões

- Objetivos
 - Identificar textos opinativos
 - Classificar o sentimento/opinião expresso no texto como positivo, negativo ou até mesmo neutro
 - Produzir um sumário claro, exibindo os resultados para o usuário
- Contexto
 - Mineração de Opiniões em *Reviews*
 - Mineração de Opiniões em Mercado Financeiro
 - Mineração de Opiniões em Debates Políticos ou Ideológicos

Como são essas opiniões?

★★★★★ **Show de bola!!!** 05/11/2014

Por: Abuhzinha

O melhor smartphone do momento.

Ótimo em todos os sentidos...

Eu já tenho um moto g agora comprei um para meu marido...

Não tenho o que reclamar dele.. excelente produto.

Quanto ao submarino, show.. estão de parabéns...

Já comprei várias vezes aqui e nunca tive problemas, entrega rápida e ótimas ofertas.

Super recomendo.

Eu recomendo este produto

Esta avaliação foi útil para você?  (1)  (0)

Como são essas opiniões?

★☆☆☆☆ **Um clichê seguido de outro** 14 de outubro de 2014

Por Inzlcs

Formato: eBook Kindle | **Compra verificada**

Estava muito interessado em ler algo da J.K. Rowling fora do universo Harry Potter, mas não fui feliz na escolha. Tenho costume de ler séries sobre detetives e esse foge de qualquer fórmula de sucesso criada por outros autores. Repetitivo, previsível e simples, o mistério a ser resolvido perde espaço para a vida do personagem principal e sua secretária (todos os outros personagens são propositalmente desprezíveis para que o leitor não tenha opção para quem torcer!).

Esta avaliação foi útil para você?

Como são essas opiniões?



GordoGeek #ForaPT @GordoGeek · 2 h

A coisa tá tão feia q até os pagamentos a **Dilma** anda atrasando. Isso aqui virou Argentina msm gordogeek.com/1EvaQI1



Com vuvuzelas, servidores do Itamaraty protestam durante almoço de ministro e embaixadores

LIBRANDE PARAGUASSU - O ESTADO DE S. PAULO
10 Novembro 2014 | 14h 35

Funcionários do Ministério das Relações Exteriores cobram pagamento do auxílio-moradia de quem trabalha no exterior, atrasado há dois meses

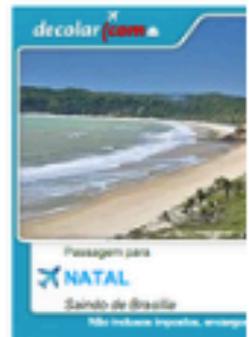
BRASÍLIA - Enquanto no segundo andar do Palácio do Itamaraty o ministro das Relações Exteriores, Luiz Alberto Figueiredo, recebia para um almoço 32

0 COMENTÁRIO(S)

DÊ A SUA OPINIÃO

SEM COMENTÁRIOS.

DÊ A SUA OPINIÃO



7



8



Como são essas opiniões?



Dilma Rousseff

20 de outubro 🌐



QUEM NÃO PULA É TUCANO

Tucano é das meias verdades. Do [#NãoVaiTerCopa](#), do País anfitrião sem aeroportos, mobilidade ou segurança. E é da amnésia do dia seguinte à [#CopaDasCopas](#)

Tucano é contra o investimento planejado e deixa o País às escuras e sem água na torneira. Que seca depois da eleição.

Tucanos inventam a tal empregabilidade e condenam 11 milhões e 400 mil ao desemprego. E ainda defendem menos trabalho e renda.

Tucano planta inflação para colher juros. E brinca com o mercado de ações, especulando, de forma completamente irresponsável.

Tucanos vêem o Brasil como uma catástrofe. Pensam o País pequeno e querem entregar a Nação gratuitamente para as grandes potências.

Como obter e analisar essas opiniões?

- *Crawlers*
- Processamento de Linguagem Natural
- Recuperação da Informação
- Mineração de Texto
- Inteligência Artificial

ENTRANDO NO UNIVERSO DE AS

Nomenclatura e definições

- Objeto
 - Um objeto é uma entidade que pode ser um produto, serviço, pessoa, evento, organização ou tópico que seja alvo de um comentário.
 - Ex.:
 - A melhor operadora de telefonia é a **VIVO**. Ela tem a maior cobertura no PI.

Nomenclatura e definições

- Atributos ou Características
 - São componentes ou aspectos de um dado objeto e que podem ser referenciados no texto.
 - Atributos explícitos, Atributos implícitos
 - ex.:
 - Apesar da bela **interface**, a **bateria** do iPhone é uma porcaria.
 - O S5 é ótimo e tem 16Gb.
 - *capacidade*

Nomenclatura e definições

- Autor

- É a pessoa ou organização que expressa a opinião. Para os exemplos acima, os autores são as pessoas que escreveram os *posts*.

- Opinião

- É uma visão, atitude, emoção ou avaliação sobre um objeto, ou sobre seus *atributos*, por parte de um autor.
- Ex.: Apesar da bela interface, a bateria do iPhone é uma porcaria.

Nomenclatura e definições

- Orientação da Opinião
 - A orientação da opinião acerca de um objeto (ou de seus atributos) pode ser positiva ,negativa ou neutra.
- Palavras Opinativas
 - São as palavras que qualificam os objetos e/ou seus atributos. Geralmente, são palavras das classes gramaticais dos adjetivos e advérbios, porém, pode-se encontrar verbos e substantivos com o sentido de qualificação

Nomenclatura e definições

- A **melhor[+]** operadora de telefonia é a **VIVO**. Ela tem a **maior[+]** cobertura no PI.
- O S5 é **ótimo[+]** e tem 16Gb.
- O *Titanic* é um **bom[+]** filme, porém o **melhor[+]** de todos os tempos é *The Godfather*.
- Apesar da **bela[-]** interface, a bateria do iPhone é uma **porcaria[-]**.
- A cerveja está **gelada[+]**.
- A comida já está **gelada[-]**.

Nomenclatura e definições

- Opiniões Diretas
 - São os casos de opiniões que fazem referência a apenas um objeto ou atributo.
- Opiniões Comparativas
 - Opiniões comparativas expressam relações de similaridade ou diferença entre dois ou mais objetos, ou atributos comuns a dois objetos, que foram destacados pelo autor. As opiniões comparativas geralmente são expressas no texto na forma de adjetivos comparativos, ou adjetivos superlativos.

Nomenclatura e definições

- Ivete Sangalo é **melhor** cantora que a Claudia Leite, só que a Cláudia é **mais nova que** a Ivete.
- O hamburguer do McDonalds é **saborosíssimo**.

Nomenclatura e definições

- Sentenças Subjetivas

- São sentenças que expressam sentimentos ou crenças pessoais

- Sentenças Objetivas

- As sentenças objetivas expressam informações factuais sobre o mundo, ou sobre o objeto sob análise
 - Ex.: **O dia amanheceu nublado.** Está bom para dormir e péssimo para ir trabalhar.

Nomenclatura e definições

- **Opinião Implícita**
 - Uma opinião implícita é aquela que está embutida em sentenças objetivas
- **Opinião Explícita**
 - Ao contrário da anterior, as opiniões explícitas são aquelas expressadas através de sentenças subjetivas
- **Sentenças Opinativas**
 - São aquelas que podem apresentar tanto opiniões explícitas como opiniões implícitas.

TRABALHANDO COM OPINIÕES

Trabalhando com opiniões

- Níveis de Granularidade de sistemas de AS
- Etapas para o processo completo de Mineração de Opiniões

Níveis de Granularidade

- Segundo Liu (2010) os níveis de granularidade podem ser:
 - Nível do Documento
 - Nível da Sentença
 - Nível da Característica/Atributo

Nível do Documento

- Busca determinar se cada documento expressa uma opinião geral positiva, negativa ou neutra.
 - A polaridade neutra é encontrada quando a quantidade de opiniões positivas é igual a quantidade de opiniões negativas
- Diversas abordagens de aprendizagem de máquina [Pang et al., 2002], [Sebastiani, 2002], [Dave et al., 2003]

Nível da Sentença

- Geralmente possui duas etapas:
 - Procura-se determinar se a sentença é subjetiva ou objetiva
 - Em caso de sentenças subjetivas, identificar se expressam opiniões positivas, negativas ou neutras
- Aprendizagem de Máquina
 - [Yu e Hatzivassiloglou, 2003]
- Conhecimento
 - [Castro, 2011]

Nível do Atributo

- Atributos das entidades são analisados individualmente
- Maior Nível de Granularidade
 - Determinar quais são os atributos de forma manual ou automática
 - Identificar e classificar as opiniões associadas a estes atributos
- [Silva et al., 2012]

Etapas da AS

- Detecção de Subjetividade
- Extração de Características/Atributos
- Classificação de Sentimento
- Apresentação dos Resultados

Etapas da AS

- Detecção de Subjetividade
- Extração de Características/Atributos
- Classificação de Sentimento
- Apresentação dos Resultados

Detecção de Subjetividade

- Detectar as sentenças subjetivas e objetivas no texto
 - As sentenças subjetivas apresentam opiniões de forma explícita
- Método Linguístico
 - Hatzivassiloglou e Wiebe (2000) defendem que adjetivos são fortes indicadores de subjetividade
 - Usam a Orientação Semântica para determinar a subjetividade

Detecção de Subjetividade

- Método Aprendizagem de Máquina
 - Yu e Hatzivassiloglou (2003) utilizaram o algoritmo Naive Bayes para classificar sentenças subjetivas/objetivas
- Atributos
 - Unigramas, classes gramaticais e polaridade das palavras

Etapas da AS

- Detecção de Subjetividade
- **Extração de Características/Atributos**
- Classificação de Sentimento
- Apresentação dos Resultados

Extração de Atributos

- Responsável por identificar e extrair dos textos disponíveis os aspectos, componentes ou características associados ao objeto sob análise.
- Essa etapa é obrigatória quando se deseja realizar uma AS no nível de atributo

Extração de Atributos

- Liu (2012) destaca que as principais pesquisas de extração de atributos foram feitas em *reviews* de produtos
- Dois tipos de *reviews*
 - Prós, contras e *review* detalhado
 - *Reviews* de formato livre

Prós, contras e *review* detalhado

Prós

“Difícil de preço! O hotel combina pró e contra”

●●●○○○ Avaliou em Julho 2, 2013

Essa avaliação foi traduzida automaticamente do inglês [O que é isso?](#)

Exibir traduções automáticas Sim Não

Traduzido por: 
Language Weaver

Positivo: moderado taxa, muito simpática equipe e boa localização. Transporte de ida e volta do Aeroporto depois da meia-noite - US 20,00 cada caminho por 1-2 pax sem nenhum problema. McDonalds com WiFi estável 200 metros 24H, loja de conveniência atrás da esquina 24H. 10 min. a pé ao longo da estrada principal do bar, jantar, área de loja. Os restaurantes locais Ocasionalmente (principalmente às quintas-feiras) colocar barracas em frente a estrada principal à direita na praia do mar e vender vários enchimento comida local: 5-6 opções para US 5,00 diversão!

Negativo: quarto é simples e bastante maltratado, kitchenette está horrível - que não se atreve a cozinhar alguma coisa. Praia e piscina estão muito longe.

Qualidade Misto de sono:

- Cama é espaçoso e moderadamente difícil, travesseiros são OK, sem o ruído do lado de fora (a minha janela ficava a entrada), lençóis limpos;
- Sem cortinas, apenas simples e parcialmente deformada plástico duro persianas e meu quarto janela ficava comum varanda (privacidade!!), A/C é um pouco barulhento.

Prós, contras e *review* detalhado

- Geralmente frases curtas
- Cada sentença trata somente de um atributo

- Abordagens
 - Aprendizagem de Máquina
 - Processamento de Linguagem Natural (PLN)

Review de formato livre

★★★★★ Show de bola!!! 05/11/2014

Por: [Abuhzinha](#)

O melhor smartphone do momento.

Ótimo em todos os sentidos...

Eu já tenho um moto g agora comprei um para meu marido...

Não tenho o que reclamar dele.. excelente produto.

Quanto ao submarino, show.. estão de parabéns...

Já comprei várias vezes aqui e nunca tive problemas, entrega rápida e ótimas ofertas.

Super recomendo.

[Eu recomendo este produto](#)

Esta avaliação foi útil para você?  (1)  (0)

Review de formato livre

- Frases completas
- Vários atributos em uma única sentença
- Abordagens
 - Silva (2013) utilizou padrões linguísticos para encontrar atributos de produtos e opiniões.

Review de formato livre

- Abordagens (cont.)
 - Hu e Liu (2004) defendem um método de dois passos:

(1) Identificar os substantivos e sintagmas nominais mais frequentes
 - Diversas pessoas falando de produtos e atributos o vocabulário tende a coincidir

Review de formato livre

- Abordagens (cont.)

- Hu e Liu (2004) defendem um método de dois passos:

(2) Identificar os atributos menos frequentes utilizando palavras opinativas:

- Uma mesma palavra opinativa pode ser utilizada para descrever diferentes atributos.
- Palavra opinativas que avaliam os atributos frequentes podem avaliar os atributos não tão frequentes,

Etapas da AS

- Detecção de Subjetividade
- Extração de Características/Atributos
- **Classificação de Sentimento**
- Apresentação dos Resultados

Classificação de Sentimento

- Principal etapa na Mineração de Opiniões
- Cada opinião possui um valor associado
 - Polaridade da opinião
 - Negativa, positiva e neutra
- Principais abordagens
 - Aprendizagem de Máquina
 - Baseado em Conhecimento
 - Ferramentas Linguísticas

Classificação de Sentimento

- Abordagem genérica de Liu (2010)
 - Identificação das palavras opinativas
 - Positivas [+1], Negativas [-1], Neutras [0]
 - Expressões negativas
 - Invertem a polaridade das opiniões
 - Uso de uma janela para determinar proximidade a opinião
 - Ex.: não, nenhum, nada, nunca, jamais, etc.

Classificação de Sentimento

- Abordagem genérica de Liu (2010) (cont.)
 - Cláusulas adversativas
 - Sentenças com essas cláusulas trazem opiniões contrárias
 - Ex.: mas, porém , contudo, no entanto, entretanto, etc.

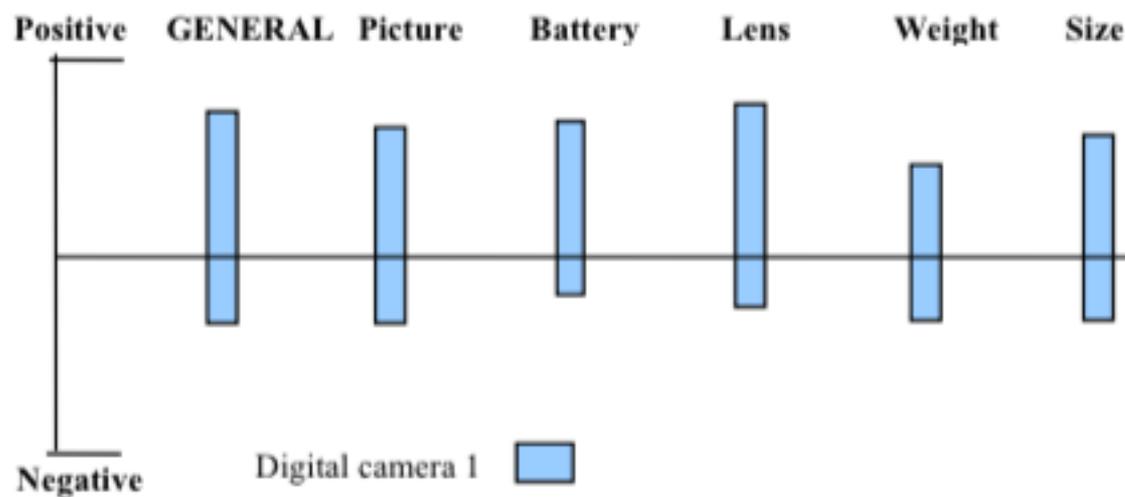
Etapas da AS

- Detecção de Subjetividade
- Extração de Características/Atributos
- Classificação de Sentimento
- **Apresentação dos Resultados**

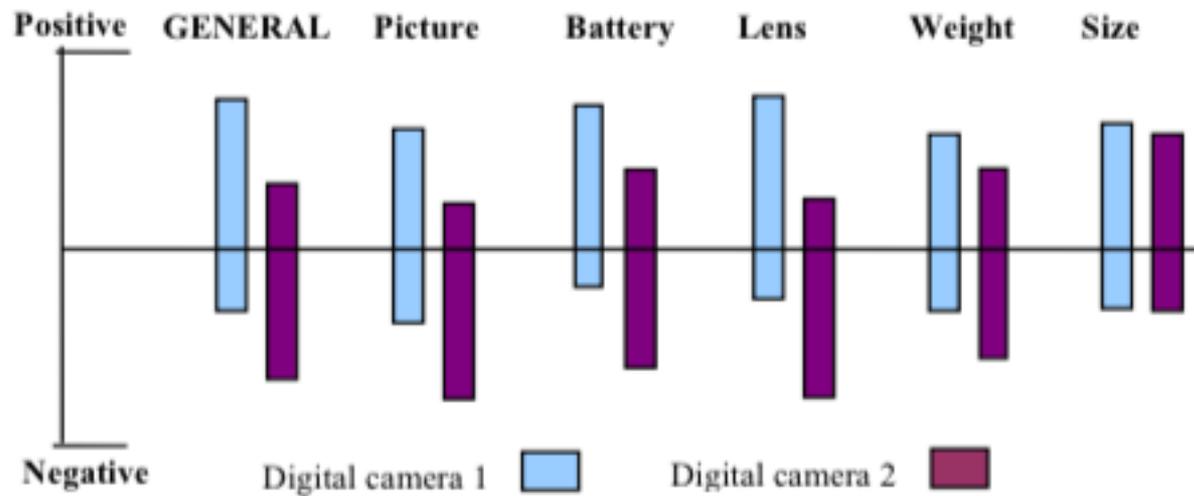
Apresentação de Resultados

- Responsável por analisar os dados de saída das etapas anteriores e apresentar ao usuário
- Pang e Lee (2008) dividem a Visualização e Sumarização em dois tipos distintos
 - Sumarização de documento único
 - entidades de um único documento
 - Sumarização de multi-documento
 - análise dos objetos sobre diversos documentos de um mesmo domínio

Sumarização de documento único



Sumarização de multi-documento



TÉCNICAS

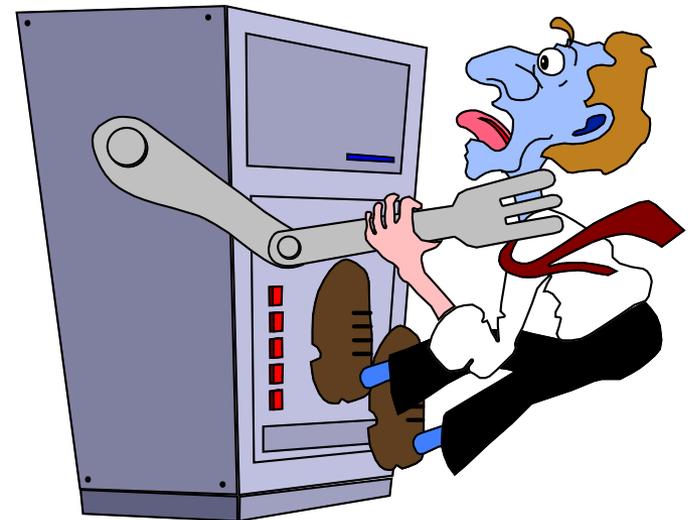
CLASSIFICAÇÃO DE TEXTOS

Roteiro

- Definição e conceitos básicos;
- Técnicas de construção de classificadores:
 - Manual;
 - Automática;
- Exemplo;
- Avaliação.

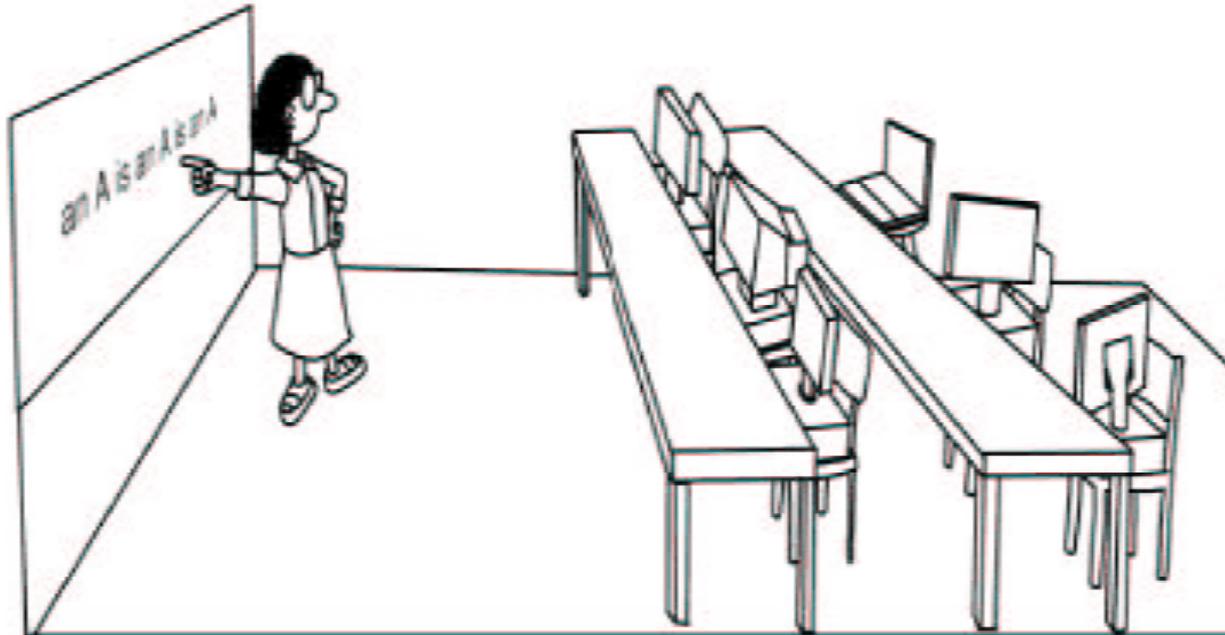
Inteligência Artificial

- Uma área de pesquisa que investiga formas de habilitar o computador a realizar tarefas nas quais, até o momento, o ser humano tem um melhor desempenho. (*Elaine Rich*)



Aprendizagem de Máquina

- O que é aprendizagem de máquina?
 - Área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais sobre o processo de aprendizado (*Bishop*)



Aprendizagem de Máquina

- Surgiu da ideia de criar programas que **aprendam** um determinado **comportamento ou padrão** automaticamente a partir de exemplos ou observações;
- Guarda alguma relação com o aprendizado humano;

Aprendizagem de Máquina

- Multidisciplinaridade:
 - Estatística;
 - Teoria da Informação;
 - Teoria do Controle;
 - Medicina
 - Filosofia;
 - Psicologia;
 - Neurobiologia;
 -

Aprendizado de Máquina

- Aprender, pode ser caracterizado como a capacidade de obter melhor desempenho pela **experiência**;

“Um programa aprende a partir da experiência E , em relação a uma classe de tarefas T , com medida de desempenho P , se seu desempenho em T , medido por P , melhora com E ”

MITCHELL, 1997

Aprendizado de Máquina

- Detecção de bons clientes para um cartão de crédito:
 - **Tarefa T**: classificar potenciais novos clientes como bons ou maus pagadores;
 - **Medida de Desempenho P**: porcentagem de clientes classificados corretamente;
 - **Experiência de Treinamento E**: uma base de dados histórica em que os clientes já conhecidos são previamente classificados como bons ou maus pagadores.

Aprendizado de Máquina

- Reconhecimento de caracteres manuscritos:
 - **Tarefa T:** reconhecer e classificar caracteres manuscritos;
 - **Medida de Desempenho P:** percentagem de caracteres classificados corretamente;
 - **Experiência de Treinamento E:** base de dados de caracteres manuscritos com a respectiva classificação.

Aprendizado de Máquina

- Produção de jogos e histórias interativas:
 - Como modelar o ambiente físico e o comportamento/personalidade dos personagens?
 - Como permitir uma boa interação com usuário?



Aprendizado de Máquina

- Controle de robôs:
 - Como obter navegação segura e eficiente, estabilidade, manipulação fina e versátil?
 - E no caso de ambientes dinâmicos e imprevisíveis?



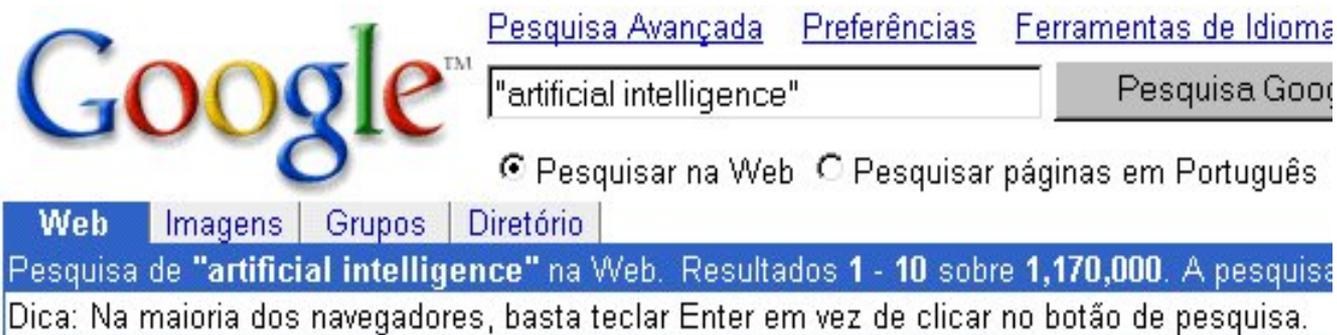
Aprendizado de Máquina

- Previsão:
 - Como prever o valor do dólar (ou o clima) amanhã?
 - Que dados são relevantes? Há comportamentos recorrentes?



Aprendizado de Máquina

- Busca de informações na Web:
 - Como localizar a informação relevante?



The image shows a screenshot of a Google search page. At the top left is the Google logo. To its right are links for "Pesquisa Avançada", "Preferências", and "Ferramentas de Idioma". Below these is a search input field containing the text "artificial intelligence" and a "Pesquisa Google" button. Under the input field are radio buttons for "Pesquisar na Web" (selected) and "Pesquisar páginas em Português". Below the search bar is a navigation bar with tabs for "Web", "Imagens", "Grupos", and "Diretório". The main content area shows the search results for "artificial intelligence" on the Web, with 1-10 results out of 1,170,000. A tip below the results reads: "Dica: Na maioria dos navegadores, basta teclar Enter em vez de clicar no botão de pesquisa."

[MIT Artificial Intelligence Lab](#) - [[Traduzir esta página](#)]

... The **Artificial Intelligence** Laboratory has been an active entity at MIT in one form or another since at least 1959. Our goal is ...

Descrição: Aiming to understand the nature of **intelligence**, to engineer systems that exhibit such **intelligence**...

Categoria:

[Computers](#) > [Artificial Intelligence](#) > [Academic Departments](#)

www.ai.mit.edu/ - 22k - [Em cache](#) - [Páginas Semelhantes](#)

Links Patrocinados

[Will robots ever think?](#)

Here are the main models of cognition, so judge for yourself.

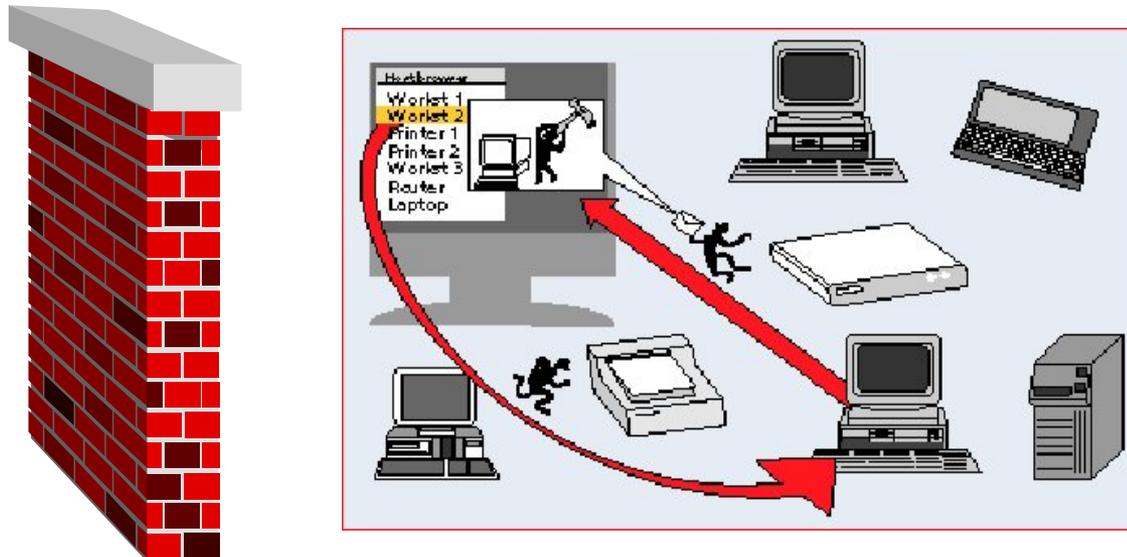
www.smithsrisca.demon.co.uk/

Interesse: 

[Veja a sua mensagem](#)

Aprendizado de Máquina

- Detecção de Intrusão e Filtragem de Spam:
 - Como saber se uma mensagem é lixo ou de fato interessa?
 - Como saber se um dado comportamento de usuário é suspeito e como lidar com isto?



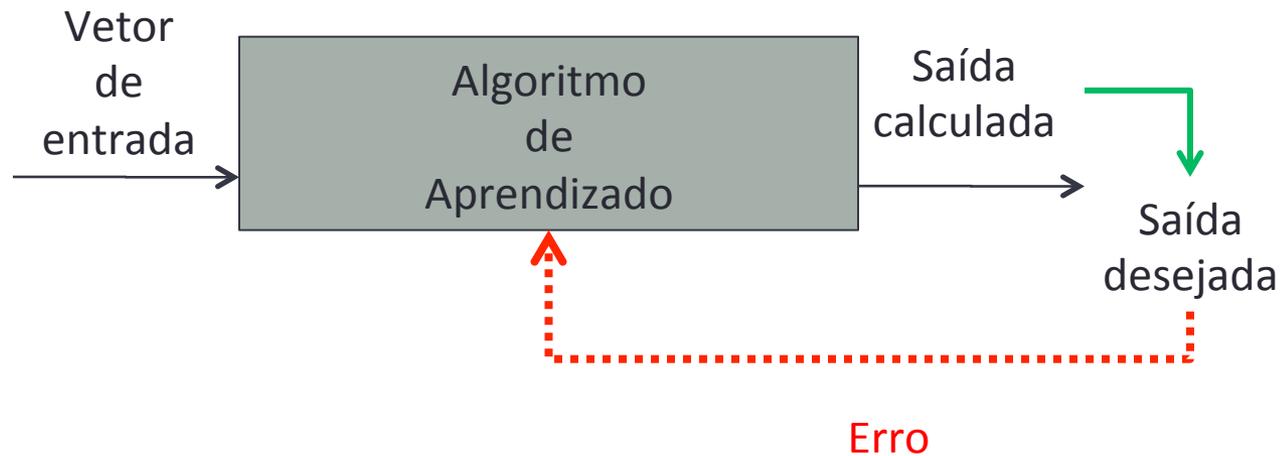
Formas de Aprendizado

- Aprendizado Supervisionado;
- Aprendizado Não Supervisionado;
- Aprendizado por Reforço.

Aprendizado Supervisionado

- É fornecida uma referência do objetivo a ser alcançado:
 - O algoritmo de aprendizado recebe o valor de saída desejado para cada conjunto de dados de entrada apresentado;
- Envolve o aprendizado de uma função a partir de exemplos de sua entrada e saída

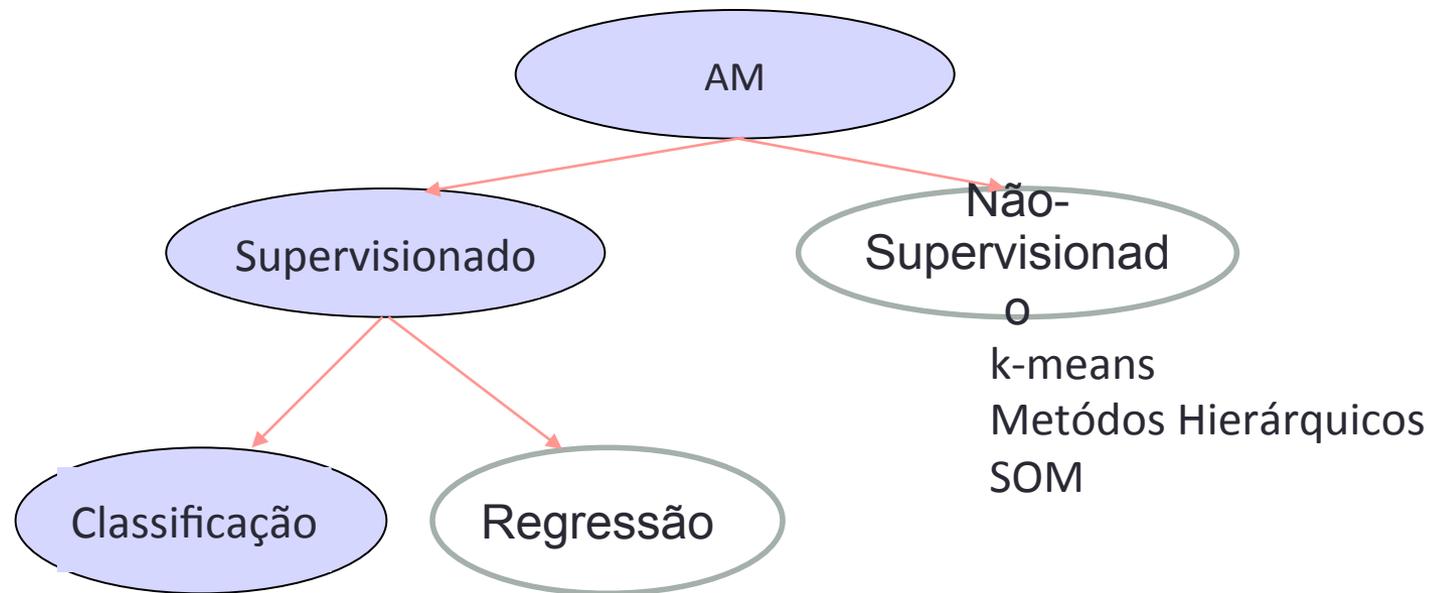
Aprendizado Supervisionado



Aprendizado Supervisionado

- Procedimento básico de treinamento:
 1. Coletar um grande conjunto de exemplos;
 2. Dividir este conjunto em dois subconjuntos distintos: conjunto de treinamento e o conjunto de testes;
 3. Treinar o algoritmo de aprendizado junto ao conjunto de treinamento;
 4. Simular o algoritmo de aprendizado treinado no conjunto de testes e medir a porcentagem de exemplos corretamente classificados;
 5. Repetir passos de 1 a 5 para diferentes tamanhos de conjuntos de treinamento e diferentes conjuntos de testes.

Aprendizagem Supervisionada



k-NN
SVM
Naive Bayes
Perceptron/Adaline
Multi-Layer Perceptron

k-NN
Adaline
Multi-Layer Perceptron

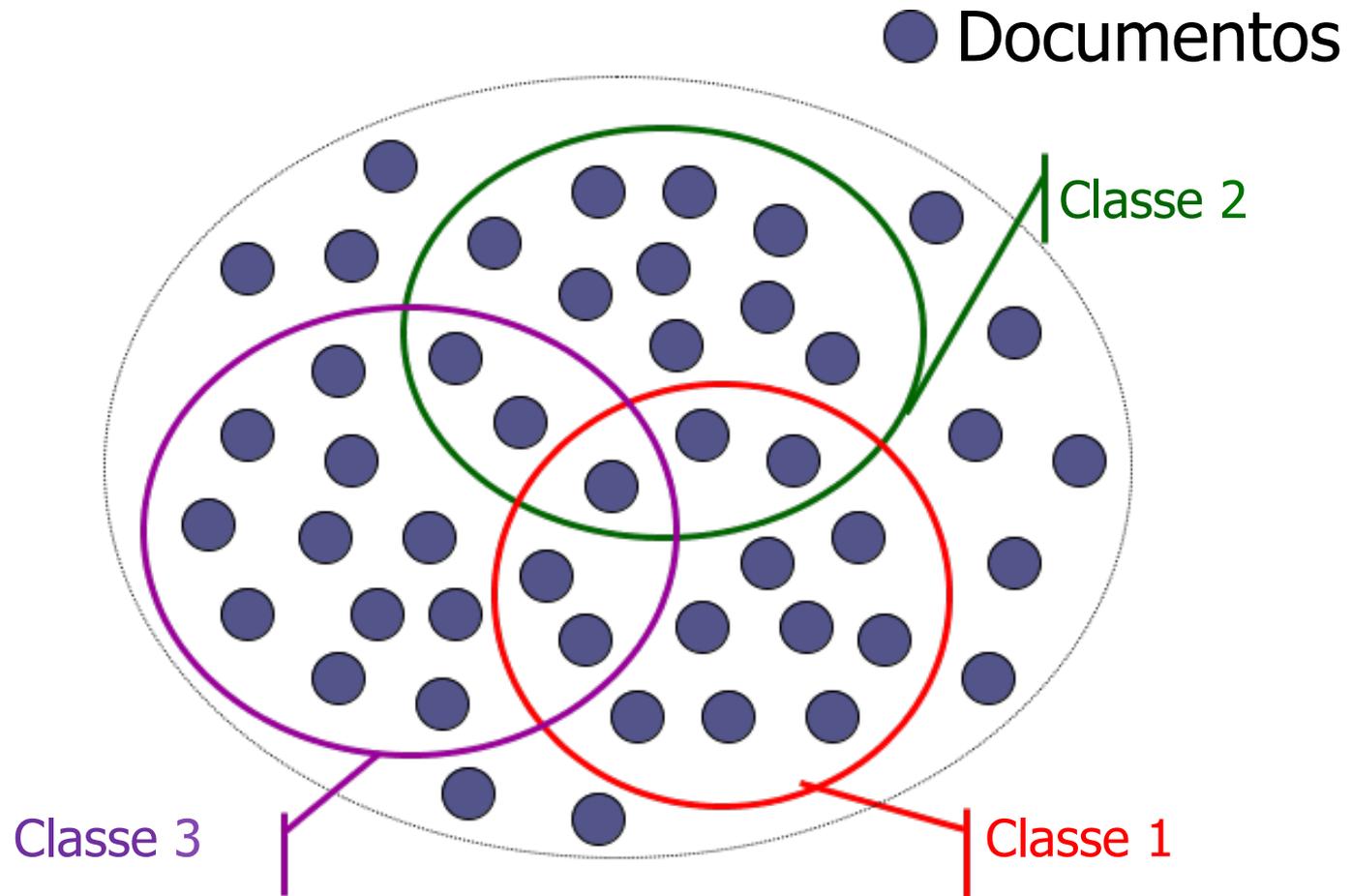
Classificação de Texto

- Objetivo:
 - Classificar documentos de acordo com **classes pré-definidas**:
 - Ex: E-mail a ser classificado como SPAM ou NÃO-SPAM;
 - Ex: Documento classificado como SENTIMENTO POSITIVO ou SENTIMENTO NEGATIVO;
 - Ex: Classificação de páginas Web (Esportiva, Política);
 - Classificação de Tweets;
 - Extração de referências bibliográficas;

Classificação de Texto

- A classificação de documentos visa:
 - Melhorar a organização da base de documentos;
 - Facilitar a busca e a visualização de documentos;
 - Ex: Hierarquias do Yahoo;
 - Evitar retornar documentos irrelevantes e/ou indesejáveis para o usuário
 -

Classificação de Texto



Classificação de Texto

- A tarefa de classificar documentos pode ser realizada de forma:
 - Manual:
 - Realizada por um especialista
 - **Automática:**
 - Realizada por um sistema automático de classificação

Classificação de Texto – Manual

- Realizada por um especialista:
 - Vantagens:
 - Alta precisão na classificação;
 - Desvantagens:
 - Tarefa demorada e trabalhosa;

Classificação de Texto – Automática

- É necessário construir um sistema que fará a classificação automática dos documentos;
 - Vantagens:
 - Rapidez na classificação;
 - Desvantagens:
 - Dificuldades na construção do sistema;
 - Perda de precisão na classificação.

Classificação Automática

- Abordagens para construção do classificador:
 - Construção manual:
 - Engenharia de Conhecimento;
 - **Construção automática:**
 - **Aprendizagem de máquina**

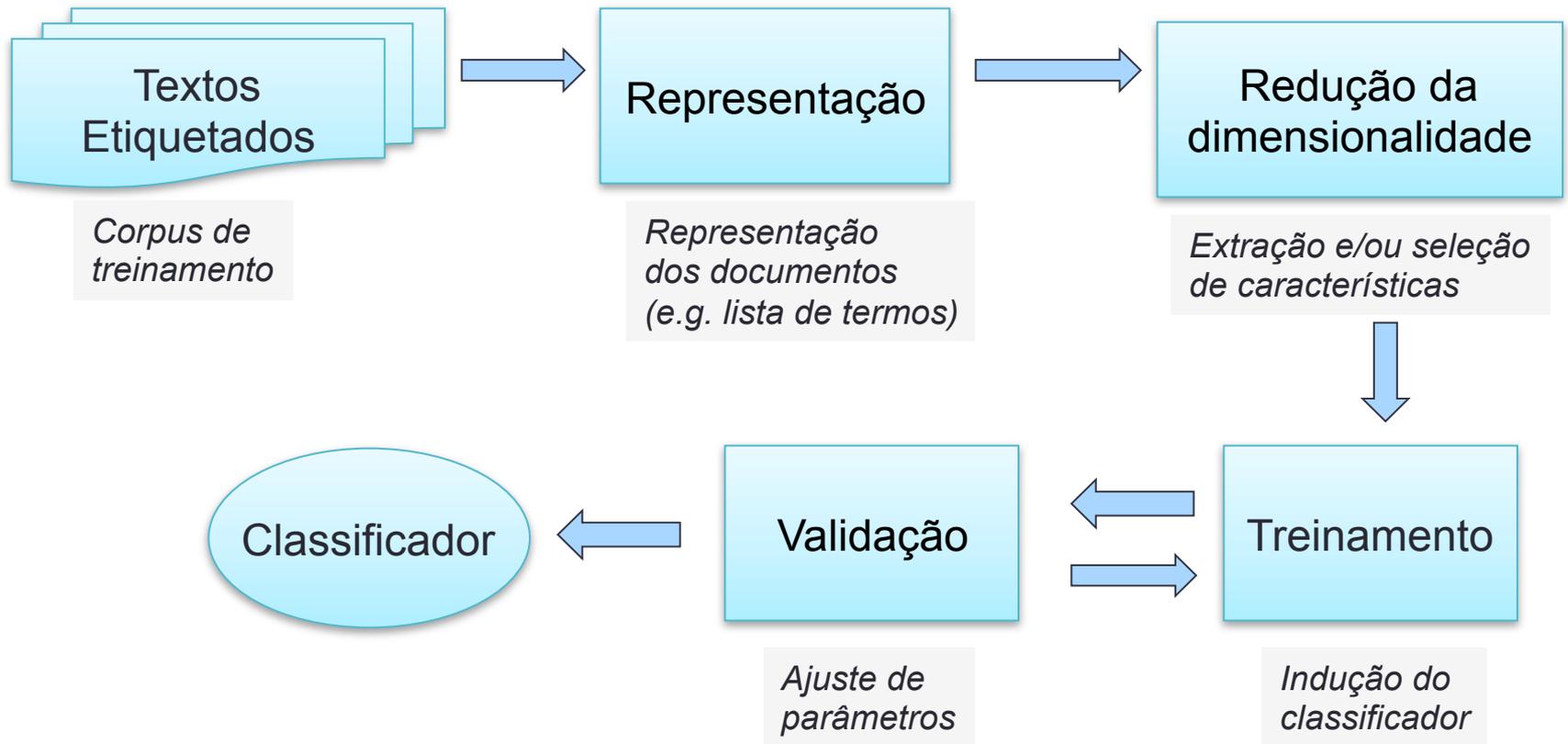
Construção Automática do Classificador

- Abordagem dominante a partir da década de 90;
- Baseada em Aprendizagem de Máquina;

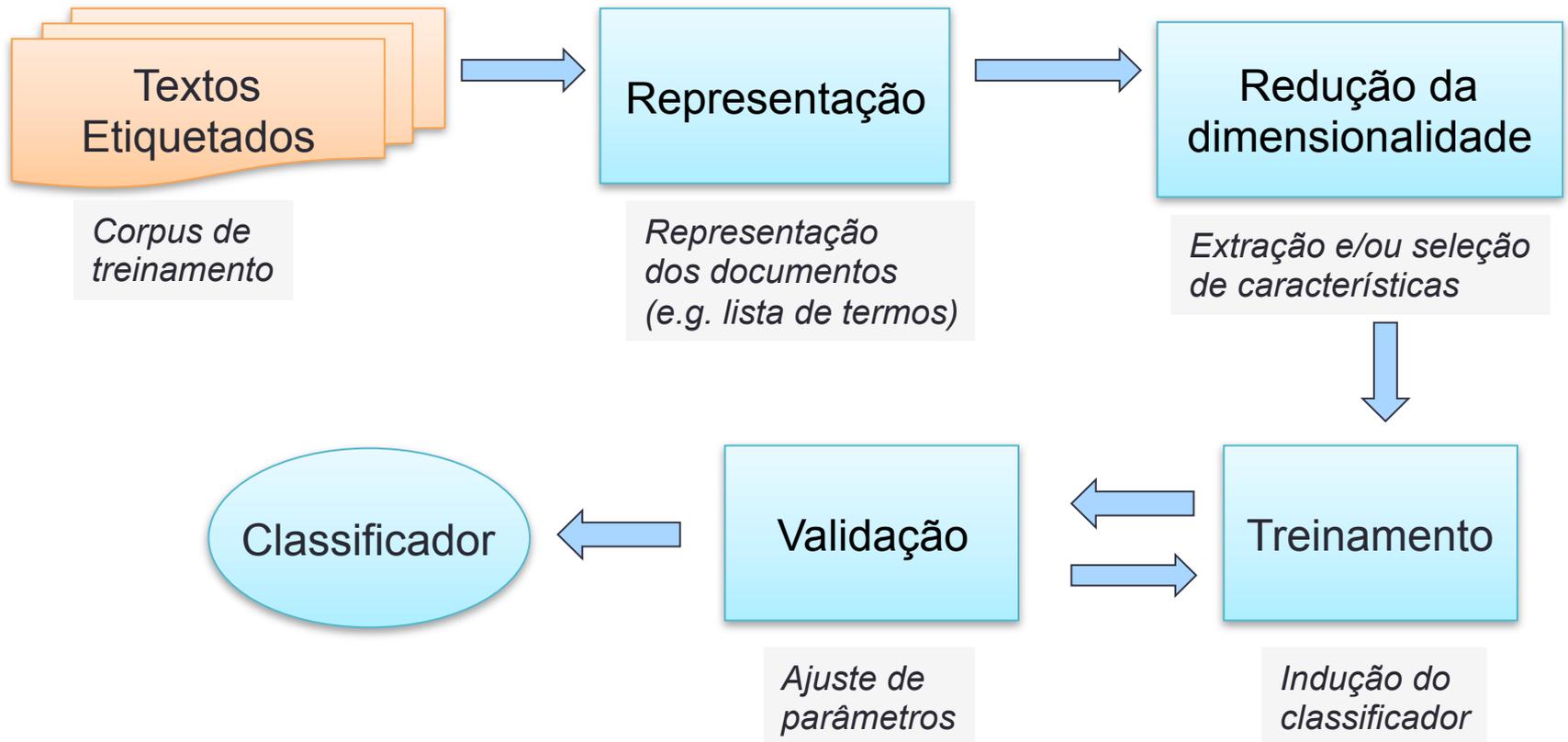
Construção Automática do Classificador

- Classes são mais simples de serem definidas através de exemplos;
- Exemplos são facilmente obtidos:
 - Especialista:
 - “Esses 20 e-mails são spam, esses 50 não”
- Necessidade de atualizar ou modificar frequentemente o classificador:
 - Usuário:
 - “Agora eu quero trabalhar no domínio de produtos eletrônicos.”
- Solução: **Aprendizagem de Máquina**

Etapas



Etapas



Processamento de Linguagem Natural

- Definição:
 - É o conjunto de métodos formais utilizados para analisar textos e extrair informações estruturadas.
- PLN normalmente faz uso de conceitos linguísticos, classes gramaticais, estrutura gramatical, além de lidar com ambiguidades e anáforas.

PLN – Corpus

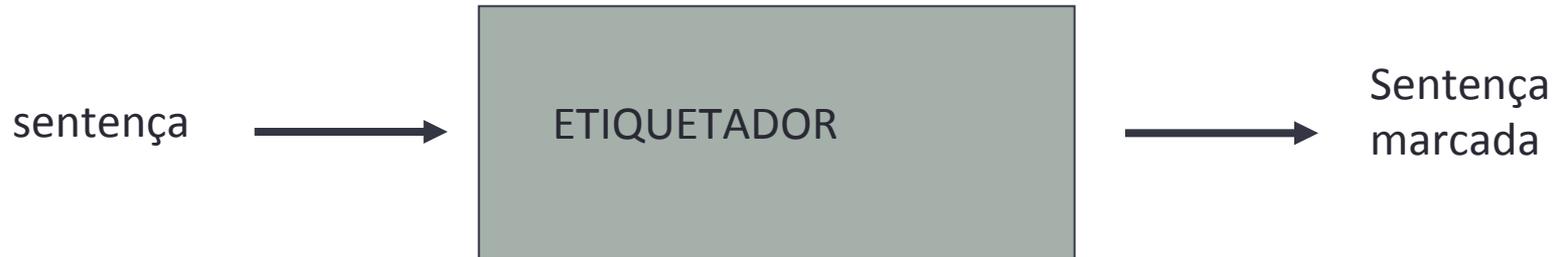
- Definição:
 - Coleção de textos:
 - Escritos ou orais;
 - Representativos da linguagem;
 - Computacionalmente disponíveis;
 - O plural – corpora
- Utilizado para treino e teste de sistemas de PLN.

Textos Etiquetados

- Etiquetadores ou *POS Taggers* ou *taggers*, são sistemas que analisam um texto e inserem etiquetas morfológicas a cada item lexical;

O iPhone é bonito			
O	iPhone	é	bonito
ARTIGO	SUBSTANTIVO	VERBO	ADJETIVO

Textos Etiquetados



Quem casa, quer casa.

Quem/pronome indef.

casa/verbo intransitivo

,/vírgula

quer/verbo trans. direto

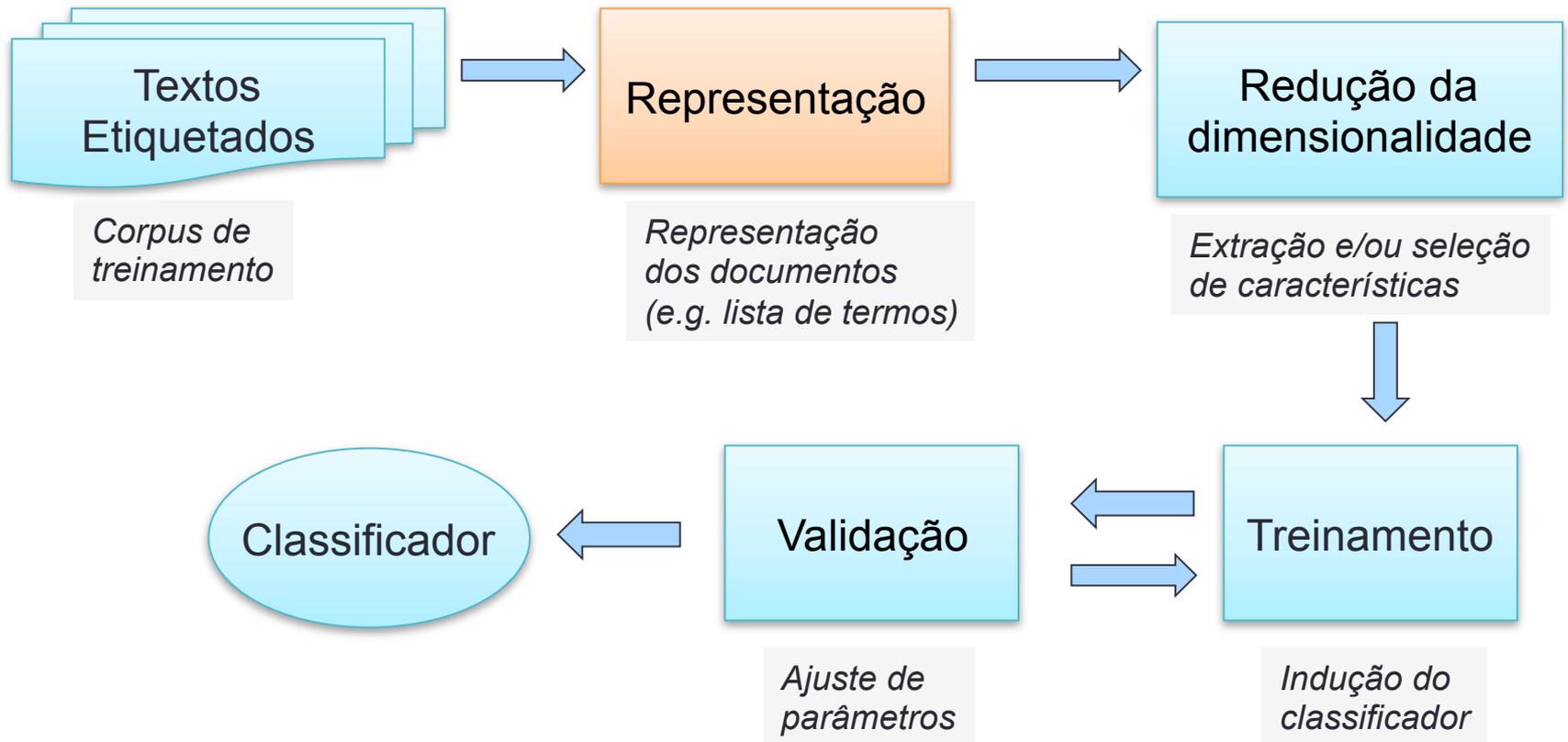
casa/substantivo comum

./ponto final

Texto Etiquetados

- Ferramentas:
 - PALAVRAS;
 - TreeTagger;
 - MXPOST;
 - Qtag;
 -

Etapas



Representação

- Seja D o conjunto de documentos e T o conjunto de termos distintos com ocorrência em D .

$$D = \{d_1, \dots, d_i, \dots, d_{|D|}\} \quad T = \{t_1, \dots, t_j, \dots, t_{|T|}\}$$

- Cada documento d em D é representado pelos termos de T associados a pesos.

$$d_i = ((t_1, w_{i,1}), \dots, (t_{|T|}, w_{i,|T|}))$$

- O peso representa a importância do termo para o documento. Quando o termo não aparece no documento, o peso associado é zero.

Representação

- Exemplo:

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

$T = \{a, \text{deixe}, e, \text{mudança}, \text{mudar}, \text{mundo}, \text{no}, o, \text{poderá}, \text{que}, \text{quer}, \text{seja}, \text{ver}, \text{você}\}$

Abordagens mais comuns para determinação dos pesos

- Booleano:
 - A representação indica apenas se o termo está ou não presente no documento;
- TF – *Text Frequency*:
 - Número de ocorrências do termo t no documento d ;
 - Tal abordagem considera que quanto mais um termo ocorre num documento, mais relevante ele é na sua representação:
 - **Nem sempre é verdade!**

Abordagens mais comuns para determinação dos pesos

- TF – IDF (*Text Frequency - Inverse Document Frequency*)
 - Reduz a importância de termos quando presentes em outros documentos;

$$TFIDF(d, t) = TF(t, d) \cdot \log \frac{|D|}{DF(t)}$$

$$D = \{d_1, \dots, d_i, \dots, d_{|D|}\}$$

$TF(t, d)$ Frequência do termo t no documento d .

$DF(t, d)$ Número de documentos em D que possuem o termo t

Representação

- Ex: **TF**

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

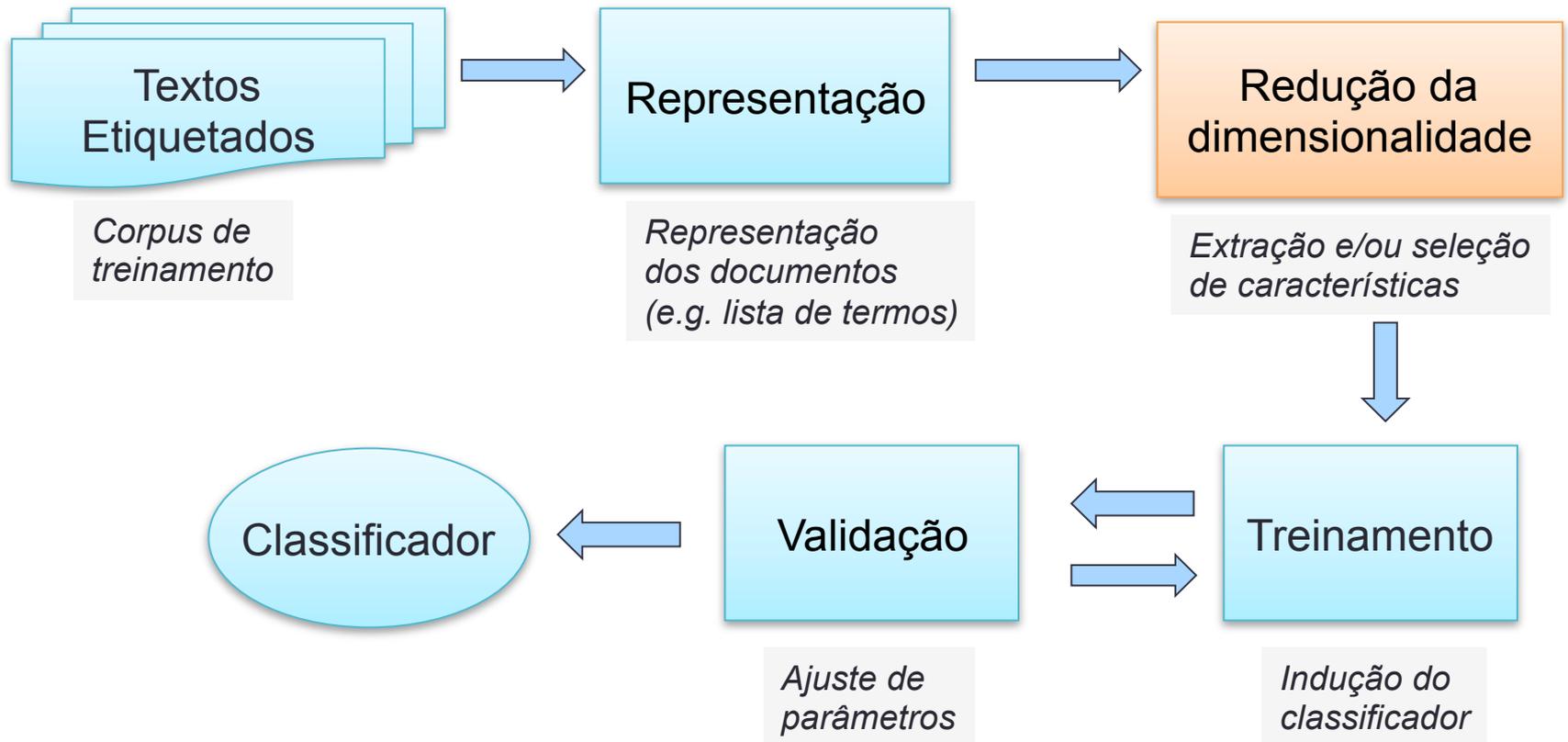
Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

T = {a, deixe, e, mudança, mudar, mundo, no, o, poderá, que, quer, seja, ver, você}

Documentos	Termos													
	a	deixe	e	mudança	mudar	mundo	no	o	poderá	que	quer	seja	ver	você
d1	1	0	0	1	0	1	1	0	0	1	1	1	1	1
d2	0	1	1	0	2	2	0	2	1	0	0	0	0	2

Etapas



Redução da Dimensão

- A redução da dimensão tem como objetivos:
 - Diminuir a complexidade do problema;
 - Eliminar termos considerados ruído na tarefa de agrupamento.
- A qualidade de um sistema de classificação de texto e agrupamento de texto são altamente **dependentes desta etapa!**

Redução da Dimensão

- Principais abordagens para redução da dimensão:
 - *Stopwords* e *Stemming*;
 - Extração de Atributos;
 - Seleção de Atributos.

Redução da Dimensão

- Principais abordagens para redução da dimensão:
 - *Stopwords e Stemming;*
 - Extração de Atributos;
 - Seleção de Atributos.

Stopwords e Stemming

- *Stopwords*:
 - Consiste na exclusão de palavras muito comuns em textos, tais como: artigos, preposições, conjunções,
- *Stemming*: (Normalização)
 - Tem como objetivo substituir a palavra por seu radical (*stem*);
 - Ex: **livro**, livrinho, livreiro, livreco, livraria;
 - Em português é bem mais complexo aplicar esse tipo de algoritmo

Stopwords e Stemming

- Ferramentas:
 - *Stopwords*: (listas genéricas);
 - <http://snowball.tartarus.org/algorithms/portuguese/stemmer.htm>
 - *Stemming*: PTStemmer
 - <http://code.google.com/p/ptstemmer/>
 - Orenge e Huyck;
 - Porter;
 - Savoy;

Stopwords e Stemming

- Ex: *stopwords* e *stemming*

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

$T = \{a, \text{deixe}, e, \text{mudança}, \text{mudar}, \text{mundo}, \text{no}, o, \text{poderá}, \text{que}, \text{quer}, \text{seja}, \text{ver}, \text{você}\}$

Documentos	Termos													
	a	deixe	e	mudança	mudar	mundo	no	o	poderá	que	quer	seja	ver	você
d1	1	0	0	1	0	1	1	0	0	1	1	1	1	1
d2	0	1	1	0	2	2	0	2	1	0	0	0	0	2

Stopwords e Stemming

- Ex: *stopwords* e *stemming*

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

T = {deixe, mud, mundo, poderá, quer, seja, ver, você}

Documentos	Termos							
	deixe	mud	mundo	poderá	quer	seja	ver	você
d1	0	1	1	0	1	1	1	1
d2	1	2	2	1	0	0	0	2

Redução da Dimensão

- Principais abordagens para redução da dimensão:
 - *Stopwords* e *Stemming*;
 - **Extração de Atributos;**
 - Seleção de Atributos.

Extração de Atributos

- É uma das etapas mais importantes em sistemas de classificação;
- Principais abordagens:
 - TF-IDF;
 - N-Gramas;
 - Estilometria;

Extração de Atributos

- N-Gramas:
 - A ideia básica é ver o texto como um fluxo de caracteres.
- Ex: Mineração:
 - Unigramas (N=1): M,i,n,e,r,a,ç,ã,o
 - Bigramas (N=2): _M,mi,in,ne,er,ra,aç,çã,ão,o_
 -

Extração de Atributos

- Objetivo:
 - Extrair as características mais representativas no documento;
 - Extrair as palavras que classificam o documento de forma mais rápida e precisa;
- Ex: (Em Análise de Sentimentos)
 - Extraí principalmente os adjetivos;

Extração de Atributos

- Estilometria:
 - A medida de estilo. É uma área de pesquisa da Linguística que usa técnicas estatísticas para analisar o estilo literário.
- Várias características estilométricas:
 - Léxicas;
 - Baseada em palavras;
 - Sintáticas;
 - Conteúdo específico.

Extração de Atributos

- Ex: Características estilométricas

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

Características Léxicas:

- Quantidade de Caracteres (C)

Documentos	Características	
	C	P
d1	42	9
d2	53	11

Baseado em Palavras:

- Quantidade de Palavras (P)

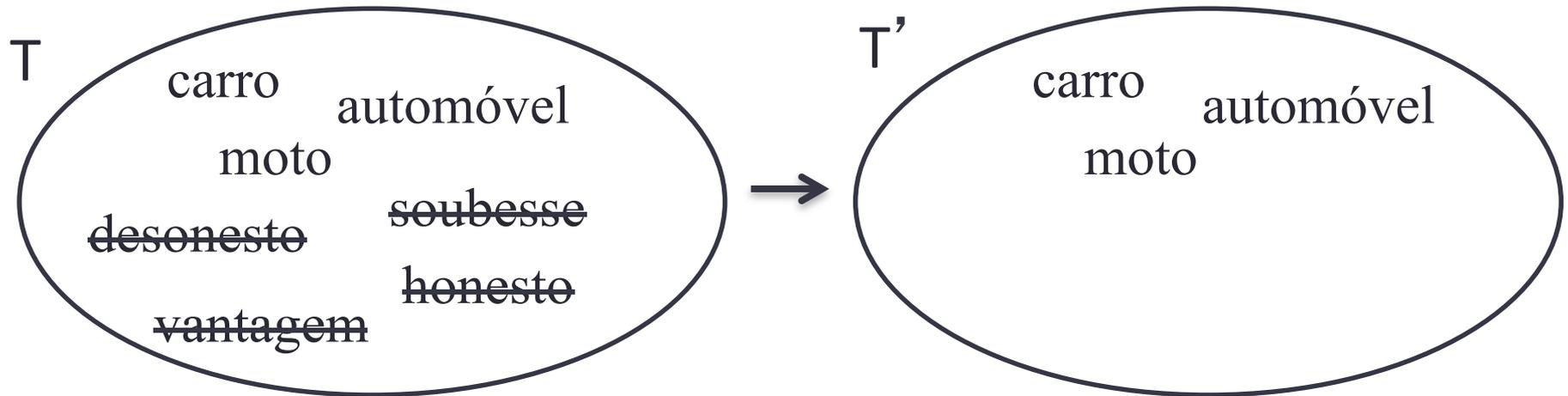
Redução da Dimensão

- Principais abordagens para redução da dimensão:
 - *Stopwords* e *Stemming*;
 - Extração de Atributos;
 - Seleção de Atributos.

Seleção de Atributos

- Seleciona os termos mais relevantes do conjunto:

$$T = \{t_1, \dots, t_j, \dots, t_{|T|}\} \longrightarrow T' = \{t_k, \dots, t_{|T'|}\}, |T'| < |T|$$



Seleção de Atributos

- Abordagens:
 - Term-Frequency;
 - Term-Frequency Inverse Document Frequency;
 - Qui Quadrado;
 - Ganho de Informação;
 -
- Algoritmos inteligentes de busca:
 - Colônia de formigas;
 - Enxame de abelhas;
 -

Seleção de Atributos

- Ex: Seleção de Atributos

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

T = {deixe, mud, mundo, poderá, quer, seja, ver, você}

Documentos	Termos							
	deixe	mud	mundo	podará	quer	seja	ver	você
d1	0	1	1	0	1	1	1	1
d2	1	2	2	1	0	0	0	2

Seleção de Atributos

- Ex: Seleção de Atributos

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

T = {deixe, mud, mundo, poderá, quer, seja, ver, você}

Documentos	Termos							
	deixe	mud	mundo	poderá	quer	seja	ver	você
d1	0	1	1	0	1	1	1	1
d2	1	2	2	1	0	0	0	2

Seleção de Atributos

- Ex: Seleção de Atributos

Texto 1:

“Seja a mudança que você quer ver no mundo.” [Dalai Lama](#)

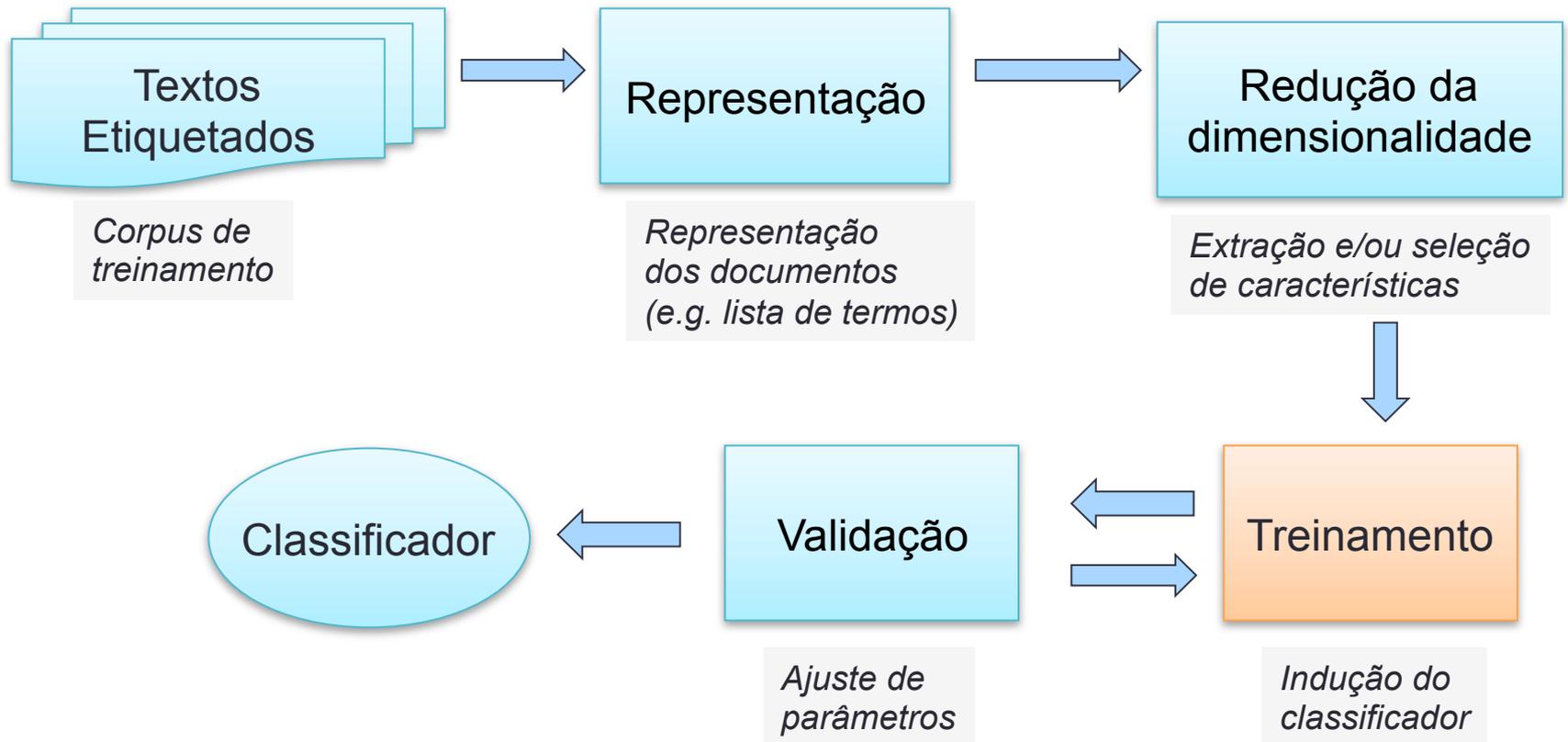
Texto 2:

“Deixe o mundo mudar você e você poderá mudar o mundo.” [Che Guevara](#)

T = {deixe, poderá, quer, seja, ver}

Documentos	Termos				
	deixe	poderá	quer	seja	ver
d1	0	0	1	1	1
d2	1	1	0	0	0

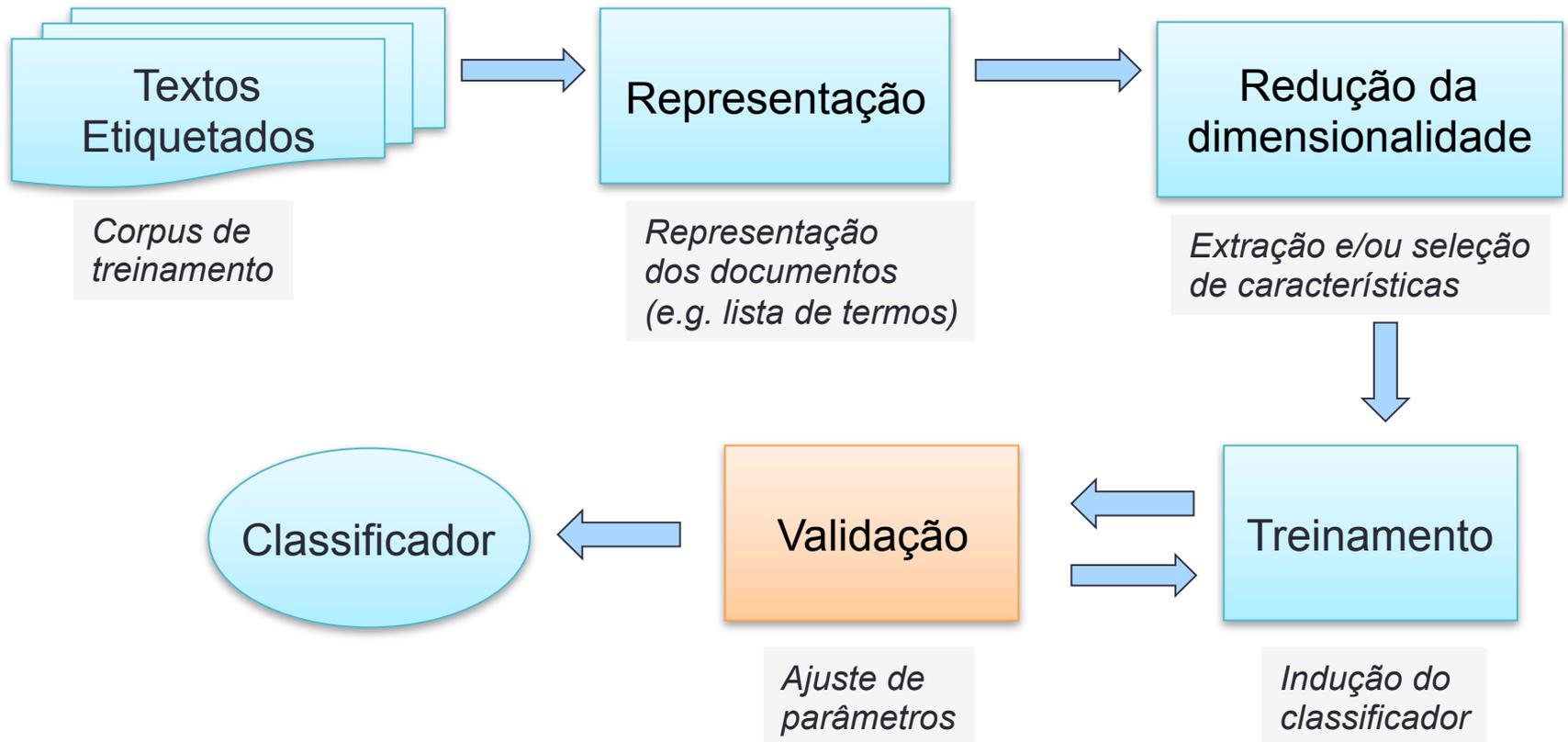
Etapas



Treinamento

- Utiliza-se um algoritmo de Aprendizagem de Máquina a partir de um *corpus*:
 - KNN;
 - Naive Bayes;
 - Árvores de Decisão;
 - Redes Neuras;
 - SVM.

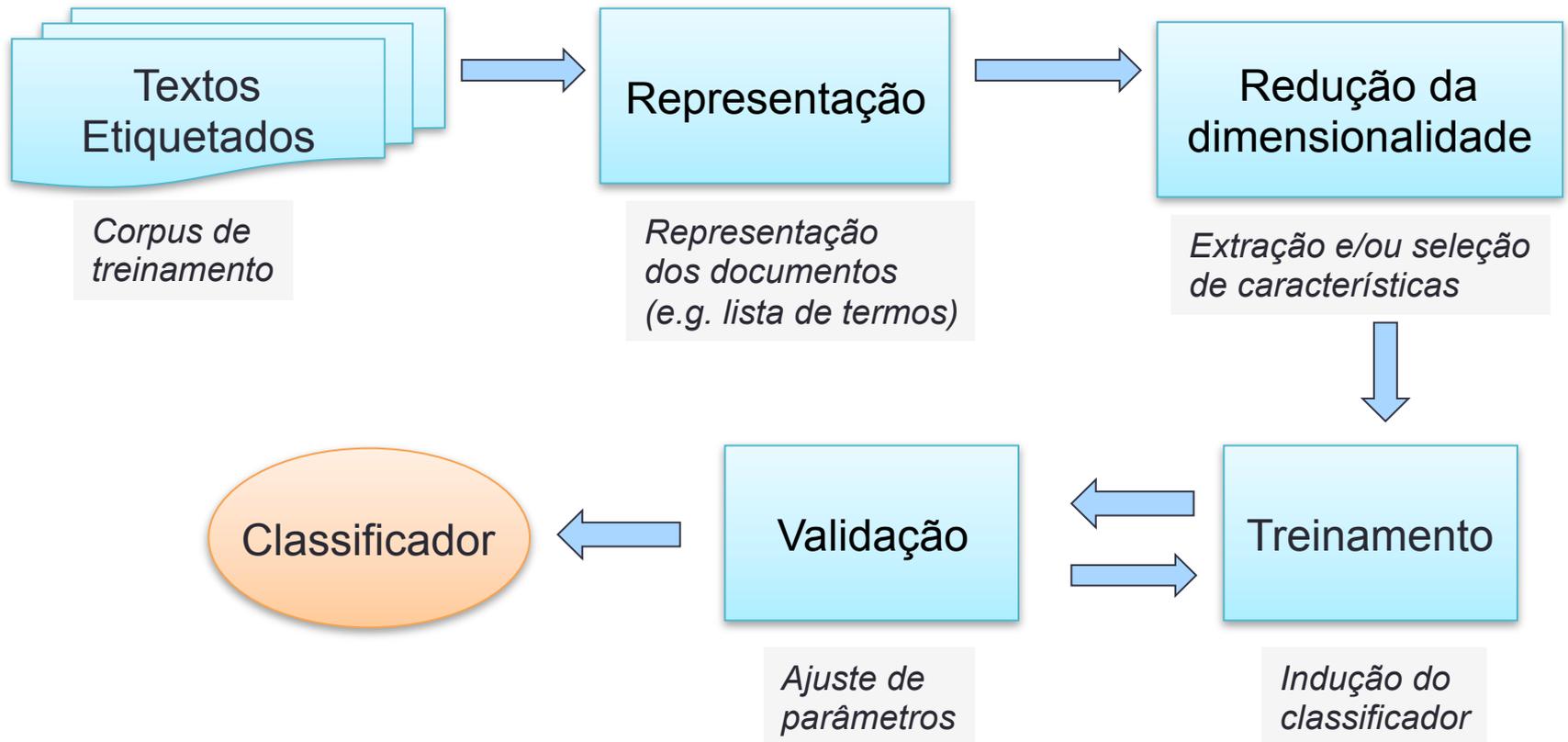
Etapas



Validação

- Validação e ajustes de parâmetros:
 - Alguns classificadores precisam de parâmetros de entrada escolhidos empiricamente:
 - Um inteiro (K) para KNN;
 - Número de nós numa Rede Neural.
 - Parte dos documentos são separados para o algoritmo automaticamente ajustar seus parâmetros, testando a precisão.

Etapas



Classificador

- Testes:
 - Por fim, os classificadores são testados:
 - Se necessário, ajustes são realizados, reiniciando o ciclo.

Usando o Naive Bayes

- Ex: Duas classes e cinco documentos de exemplo

Documento	Classe
Ruby é uma linguagem dinâmica	Linguagem
Ruby tem a cor entre o rosa e o vermelho	Pedra
Jóia (Gems) é um pacote da linguagem ruby	Linguagem
Ruby é uma das quatro pedras preciosas	Pedra
A linguagem ruby nasceu no Japão	Linguagem

Usando o Naive Bayes

- Seleccionando os atributos:

Documento	Classe
Ruby é uma linguagem dinâmica	Linguagem
Ruby tem a cor entre o rosa e o vermelho	Pedra
(Gems) Jóia é um pacote do ruby	Linguagem
Ruby é uma das quatro pedras preciosas	Pedra
A linguagem ruby nasceu no Japão	Linguagem

Usando o Naive Bayes

Termo	Linguagem	Pedra
Ruby	3	2
Jóia	1	1
Dinâmica	1	0
Linguagem	2	0
Cor	0	1
Rosa	0	1
Vermelho	0	1
Pacote	1	0
Preciosa	0	1
Japão	1	0

Usando o Naive Bayes

$$P(w_k | c_i) = \frac{(\text{valor inicial} + \text{ocorrencias de } w_k \text{ em } c_i)}{(\text{peso do valor inicial} + \text{numero de documentos em } c_i)}$$

$$P(wk | ci) = \frac{(0,5 + \text{ocorrências} - de - wk - em - ci)}{(1 + n^{\circ} - \text{documentos} - em - ci)}$$

Usando o Naive Bayes

Termo	Linguagem	Pedra
Ruby	$(0,5 + 3) / (1 + 3)$	$(0,5 + 2) / (1 + 2)$
Jóia	$(0,5 + 1) / (1 + 3)$	$(0,5 + 1) / (1 + 2)$
Dinâmica	$(0,5 + 1) / (1 + 3)$	$(0,5 + 0) / (1 + 2)$
Linguagem	$(0,5 + 2) / (1 + 3)$	$(0,5 + 0) / (1 + 2)$
Cor	$(0,5 + 0) / (1 + 3)$	$(0,5 + 1) / (1 + 2)$
Rosa	$(0,5 + 0) / (1 + 3)$	$(0,5 + 1) / (1 + 2)$
Vermelho	$(0,5 + 0) / (1 + 3)$	$(0,5 + 1) / (1 + 2)$
Pacote	$(0,5 + 1) / (1 + 3)$	$(0,5 + 0) / (1 + 2)$
Preciosa	$(0,5 + 0) / (1 + 3)$	$(0,5 + 1) / (1 + 2)$
Japão	$(0,5 + 1) / (1 + 3)$	$(0,5 + 0) / (1 + 2)$

Usando o Naive Bayes

Termo	Linguagem	Pedra
Ruby	0,88	0,83
Jóia	0,38	0,5
Dinâmica	0,38	0,17
Linguagem	0,62	0,17
Cor	0,12	0,5
Rosa	0,12	0,5
Vermelho	0,12	0,5
Pacote	0,38	0,17
Preciosa	0,12	0,5
Japão	0,38	0,17

Usando o Naive Bayes

- Ex: A qual classe pertence o documento:

“Minha jóia ruby chegou em um pacote rosa”

Usando o Naive Bayes

- Seleccionando os atributos:

“Minha **jóia ruby** chegou em um **pacote rosa**”

Usando o Naive Bayes

$$P(\text{linguagem} | d) = P(c_i) * P(d | c_i) / P(d)$$

documento em linguagem/
total de documentos = 0,6

mantém

$$P(\text{ruby} | \text{linguagem}) * P(\text{jóia} | \text{linguagem}) * \\ P(\text{rosa} | \text{linguagem}) * P(\text{pacote} | \text{linguagem})$$

Usando o Naive Bayes

$$P(\text{pedra} \mid d) = P(\text{ci}) * P(d \mid \text{ci}) / P(d)$$

documentos em pedra/
total de documentos = 0,4

mantém

$P(\text{ruby} \mid \text{pedra}) * P(\text{jóia} \mid \text{pedra}) *$
 $P(\text{rosa} \mid \text{pedra}) * P(\text{pacote} \mid \text{pedra})$

Usando o Naive Bayes

$$P(\text{linguagem} \mid d) = 0,6 * 0,88 * 0,38 * 0,12 * 0,38 / P(d)$$

$$P(\text{pedra} \mid d) = 0,4 * 0,83 * 0,5 * 0,5 * 0,17 / P(d)$$

$$P(\text{linguagem} \mid d) = 0,009149 / P(d)$$

$$P(\text{pedra} \mid d) = 0,014110 / P(d)$$

$$**P(\text{pedra} \mid d) > P(\text{linguagem} \mid d)**$$

Avaliando Classificadores

- Nomenclaturas:
 - $Tp = \textit{True positive}$:
 - Número de exemplos corretamente classificados como positivos;
 - $Fp = \textit{False positive}$:
 - Número de exemplos erroneamente classificados como positivos;
 - $Tn = \textit{True negative}$:
 - Número de exemplos corretamente classificados como negativos;
 - $Fn = \textit{False negative}$:
 - Número de exemplos erroneamente classificados como negativos;

Avaliando Classificadores

- Medidas (para uma classe):
 - Precisão e cobertura;
 - Taxa de erro;
 - Acurácia.
- A média dá a efetividade geral do classificador;

Avaliando Classificadores

- Precisão para classe C_i
 - Porcentagem de exemplos corretamente classificados como positivos:
 - $Tp / (Tp + Fp)$
- Cobertura para classe C_i
 - Porcentagem de exemplos corretamente classificados como positivos em relação ao total de instâncias da classe:
 - $Tp / (Tp + Fn)$

Avaliando Classificadores

- # total de exemplos é dado por:
 - $N = T_p + T_n + F_p + F_n$
- Acurácia:
 - Número de exemplos classificados corretamente / total de exemplos da amostra
 - $(T_p + T_n) / N$
- Erro total:
 - $(F_p + F_n) / N$

DESAFIOS

Desafios

- Reconhecimento de Entidades Nomeadas
 - *Named-Entity Recognition*(NER)
 - Reconhecimento de entidades que estão sendo avaliadas no texto
 - Dificuldade
 - Diversas entidades presentes em um só texto
 - Anáforas

Desafios

- Dicionário de palavras opinativas
 - Falsos *hits* com as palavras do dicionário
 - Conotações e polaridades diferentes em contextos distintos
- Detecção de Subjetividade
 - Determinar quais documentos possuem conteúdo opinativo
 - Ironias, sarcasmos, metáforas

Desafios

- Textos opinativos com ruídos
 - Erros de grafia
 - Erros gramaticais
 - Falta de pontuação
 - Gírias
 - Solução?
 - Pré-processamento refinado

PRÁTICA

Prática

- O que usaremos?
 - Weka
 - Eclipse / Netbeans
 - Stanford *Pos-Tagger*

Obrigado

DÚVIDAS????