



Machine Learning em Delphi - Um exemplo de classificação de textos

Rodolpho Nascimento

Da sala de reunião à vida real

Embarcadero Conference 2019

Resumo

- Introdução
- O que é Machine Learning?
- Por que o Delphi não é tão popular nas comunidades de ML?
- Tipos de aprendizado em ML
- Manipulando textos em ML - BOW
- Identificando o grau de similaridade entre textos
- Demonstração (código-fonte)
- Exemplos de uso no mercado
- Perguntas?

Rodolpho Nascimento

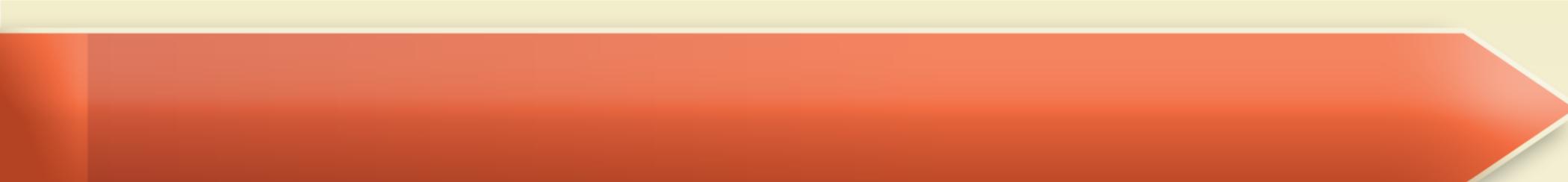
Rio de Janeiro-RJ

- Mestrando em Ciência da Computação - CEFET/RJ
 - Linha de pesquisa: Computação Afetiva
 - Especialidade: Mineração de Textos
- Formação: Análise e desenvolvimento de sistemas (UNESA)
- Trabalhando com Delphi desde a versão 5.
- Coordenador de Produtos na GKO Informática (RJ)
- Além de Delphi, desenvolvo programas em outras linguagens: C#, Typescript, Javascript, ActionScript, Dart, Python, R, etc...

Introdução

Introdução

- Objetivo desta palestra é aplicar uma técnica simples de ML puramente em Delphi
- Será abordada de forma simplista alguns conceitos teóricos
- Uma demonstração será apresentada, com análise de código-fonte
- Alguns exemplos de aplicação de mercado serão apresentados



O que é Machine Learning?

O que é Machine Learning?

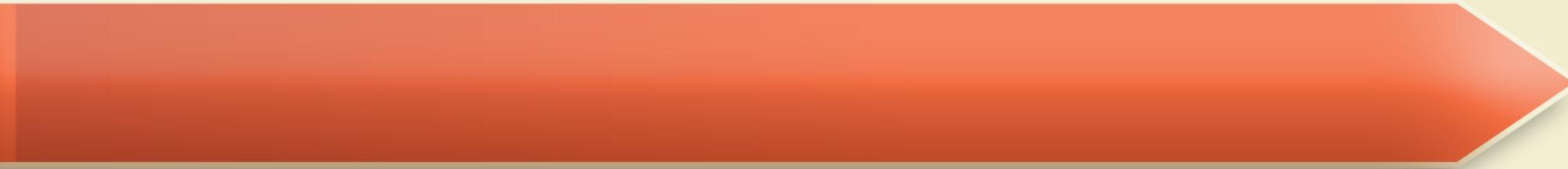
Definição de Machine Learning (ou Aprendizado de Máquina):

“Um campo de estudo que capacita os computadores para a habilidade de aprender sem serem explicitamente programados”

Arthur Samuel, “Some studies in machine learning using the game of checkers. IBM J Res Dev . 1959;3:535–554”

O que é Machine Learning?

- ML não é algo novo. Alan Turing foi um dos pioneiros dos estudos de ML publicando em 1950 o artigo COMPUTING MACHINERY AND INTELLIGENCE, com o famoso “O jogo da imitação” (*the imitation game*).
- O primeiro modelo de neurônio artificial foi proposto pelos pesquisadores Warren McCulloch e Walter Pitts, em 1943, no artigo *A Logical Calculus of the Ideas Immanent in Nervous Activity*.
- O atual momento (grande quantidade de dados + grande poder de processamento) permitiu a popularização da aplicação de ML.



Por que o Delphi não é tão popular nas
comunidades de ML?



Por que o Delphi não é popular em ML?

- É a matemática que faz a grande “mágica” em ML
- Algoritmos, são algoritmos...
- Linguagens como Python e R ganharam grande adesão de desenvolvedores na implementação de algoritmos de ML:
 - Scikit Learn (Python)
 - Gensim (Python)
 - Tensorflow (escrita inicialmente em C++, mas integra com outras linguagens)
- Em contrapartida, poucas bibliotecas de ML para Delphi
 - Mitov software (proprietário)
 - Tensowflow4Delphi (projeto no Github)

Tipos de aprendizado em ML

Tipos de aprendizado em ML

- **Aprendizado supervisionado:** Quando se fornece as respostas ao algoritmo para a execução da tarefa. Alguns algoritmos:
 - Árvore de decisão
 - Naive Bayes
 - Regressão logística
 - Regressão linear
 - Redes neurais
- **Aprendizado não-supervisionado:** O algoritmo se encarrega de executar a sua tarefa sem alguma resposta fornecida. Alguns algoritmos:
 - K-nearest neighbors (KNN)
 - K-means
 - DBSCAN

Manipulando textos em ML - BOW

Manipulando textos em ML - BOW

- Em sistemas de recuperação da informação, é necessário representar um texto de forma que facilite as operações solicitadas ao computador
- Um modelo amplamente adotado, é o Modelo Espaço Vetorial (MEV), também conhecido como Bag Of Words (BOW)
- A ordem dos termos é descartada neste modelo
- A representação um texto é dada através de um vetor com pesos (características), ex:

Texto: O gato e o rato são inimigos

BOW: [2, 1, 1, 1, 1, 1]

Manipulando textos em ML - BOW

- Uma das formas de atribuir pesos ao vetor, é a contagem de frequência de termos (Term Frequency - TF)
- A frequência do termo no texto é anotada, e atribuída ao vetor como peso, ex:

o	gato	e	o	rato	são	inimigos
2	1	1		1	1	1

$$V = [2, 1, 1, 1, 1, 1]$$

- Desta forma, já podemos aplicar algumas operações com os textos



Identificando o grau de similaridade entre textos

Identificando o grau de similaridade - textos

- Em tarefas de classificação de textos utilizando ML, é importante detectar o quanto similar é um conteúdo entre dois textos
- Como os textos estarão representados no modelo BOW, podemos utilizar uma função matemática detectar o grau de similaridade
- Dentre várias funções disponíveis, uma bastante utilizada é a função **cosseno** para calcular a **similaridade** entre os vetores

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}},$$

Identificando o grau de similaridade - textos

Texto A

o	gato	e	o	rato	são	inimigos
2	1	1		1	1	1

Multiplicar os pesos entre os vetores, que assim colidirem

Texto B

o	gato	não	gosta	do	rato
1	1	1		1	1

Resultado:

2	1	1
---	---	---

= 50% de similaridade

Identificando o grau de similaridade - textos

Texto A

o	gato	e	o	rato	são	inimigos
2	1	1		X	1	1

Multiplicar os pesos entre os vetores, que assim colidirem

Texto B

o	gato	não	gosta	do	rato
1	1	1		1	1

Resultado:

2	1	
---	---	--

= 28% de similaridade

Identificando o grau de similaridade - textos

Texto A

o	gato	e	o	rato	são	inimigos
2	1	1		1	1	1

Multiplicar os pesos entre os vetores, que assim colidirem

Texto B

o	gato	não	gosta	do	rato	são
1	1	1		1	1	1

Resultado:

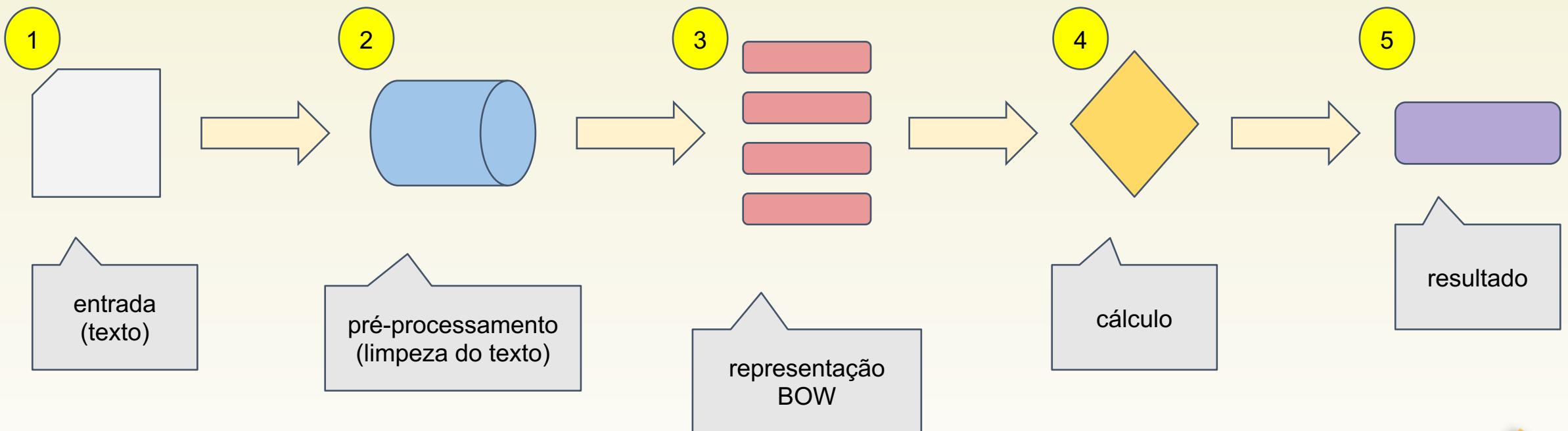
2	1	1	1
---	---	---	---

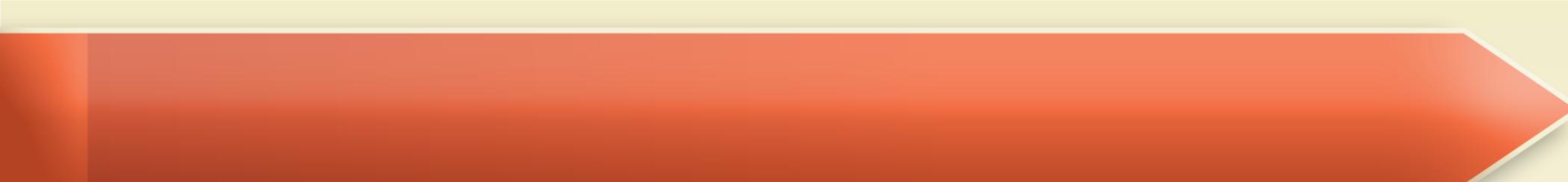
 = **67%** de similaridade

Demonstração (código-fonte)

Demonstração (código-fonte)

Visão geral





Exemplos de uso no mercado

Cenário 1 - Agência de marketing

- Uma agência de marketing foi contratada para uma publicidade esportiva
- O seu cliente precisa montar uma estratégia com base em noticiários de sites esportivos
- O sistema coleta diversas notícias dos sites e faz a classificação do conteúdo desejado, separando os artigos com o contexto mais similar

Cenário 2 - Identificar atendimentos SAC

- Um sistema de SAC faz o registros de atendimentos aos clientes de uma empresa
- Em um dos atendimentos, o cliente faz uma reclamação de um produto
- Como tarefa, o sistema deve detectar todas as reclamações registradas referentes ao produto

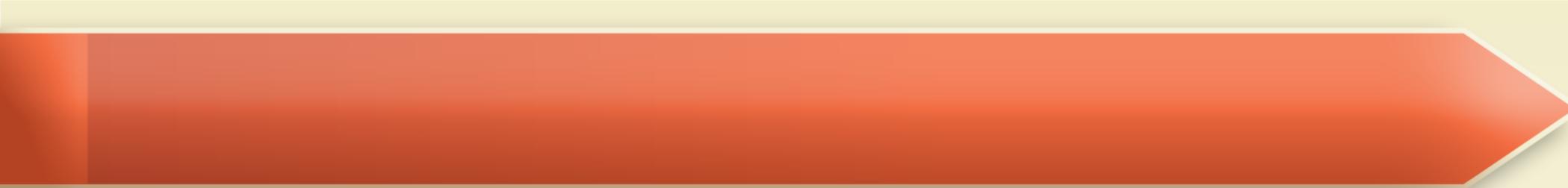
Cenário 3 - Chatbot

- Uma empresa prestadora de serviços telefônicos precisa melhorar o atendimento aos clientes, agilizando atividades rotineiras
- Uma forma de agilizar o atendimento é através da automatização de processos
- Um chatbot é disponibilizado aos clientes para detectar os pedidos mais comuns, e realizar a atividade solicitada
- A automatização permite ampliar a capacidade de atendimento e reduzir custos

Código-fonte do projeto

Link do projeto:

https://github.com/rodolphonascimento/Delphi_Doc_Similarity



Perguntas?





Obrigado



rodolpho@essencialcode.com.br

Da sala de reunião à vida real

Embarcadero Conference 2019