Introdução ao PLN com NLTK e Spacy

Semana Integrada PUCC 2019

MsC Alessandro Bokan, PhD Fernando Nóbrega

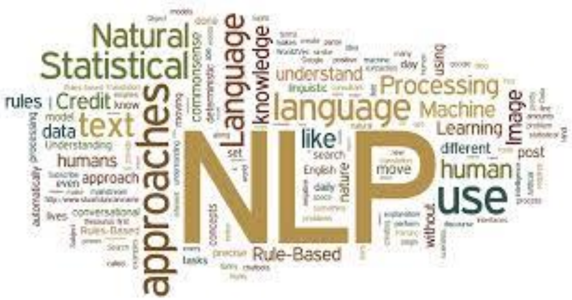
Setembro 2019

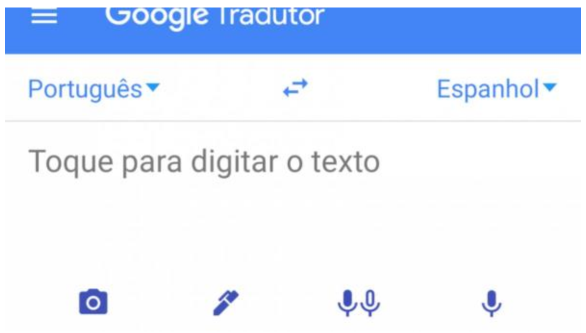
#O que é PLN?

- Processamento de Língua Natural
  - Qualquer tarefa em que a máquina deve **Interpretar** ou **Processar/Gerar** conteúdo em linguagem humana

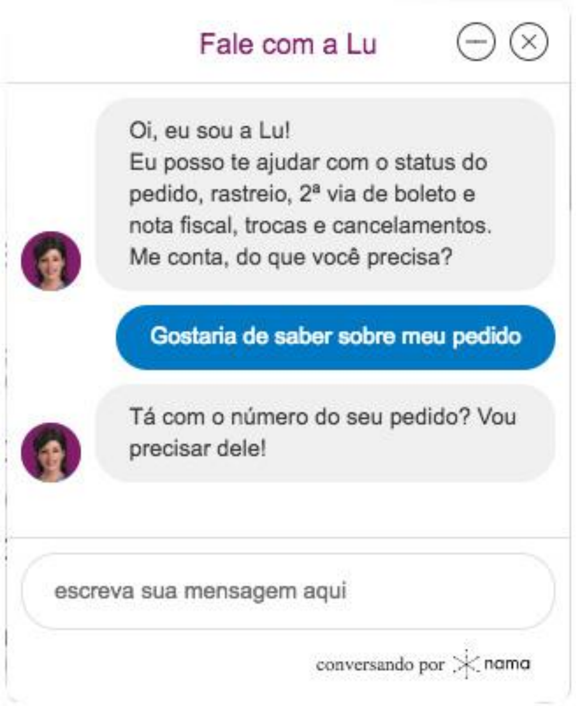
- Exemplos
  - Auto correção do celular ou word (para aqueles que gostam ou não)
  - Tradução Automática
  - Busca de Informação (Google, Bing, Baidu...)
  - Assistentes Virtuais (como o **Bixby, Siri, Alexa, Google Assistent**)
  - ChatBots
  - Mineração de Opinião
  - Geração automática de resumos
  - Entre várias outras

#O que é PLN?

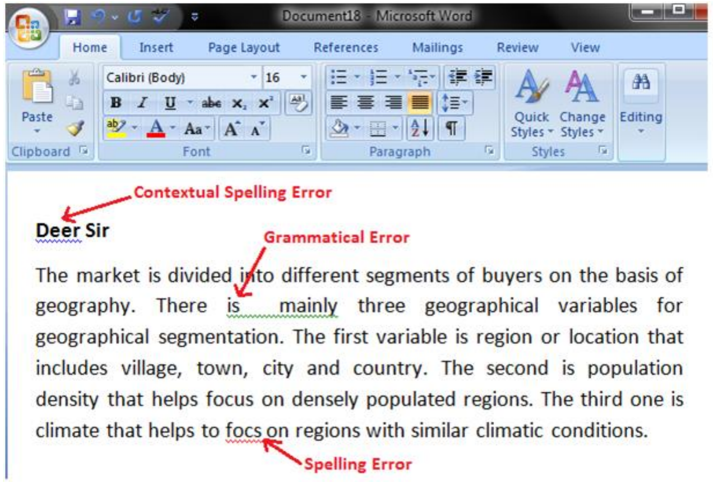

Natural
Statistical
Language
Processing
language
Machine
Learning
NLP
text
data
rules
Credit
approaches
humans
approach
human
use
like
understand
Rule-Based
knowledge


Google Tradutor
Português
Espanhol
Toque para digitar o texto


Bixby


Fale com a Lu
Oi, eu sou a Lu!
Eu posso te ajudar com o status do pedido, rastreio, 2ª via de boleto e nota fiscal, trocas e cancelamentos.
Me conta, do que você precisa?
Gostaria de saber sobre meu pedido
Tá com o número do seu pedido? Vou precisar dele!
escreva sua mensagem aqui
conversando por nama


My library
My Citations
Alerts
Metrics
Settings
Google Scholar
Articles (include patents)
Case law
Stand on the shoulders of giants


Document18 - Microsoft Word
Home
Insert
Page Layout
References
Mailings
Review
View
Paste
Clipboard
Font
Paragraph
Quick Styles
Change Styles
Styles
Editing
Contextual Spelling Error
Deer Sir
Grammatical Error
The market is divided into different segments of buyers on the basis of geography. There is mainly three geographical variables for geographical segmentation. The first variable is region or location that includes village, town, city and country. The second is population density that helps focus on densely populated regions. The third one is climate that helps to focs on regions with similar climatic conditions.
Spelling Error

#Níveis de Processamento

- Cada aplicação requer um nível diferente de "conhecimento" da línguagem

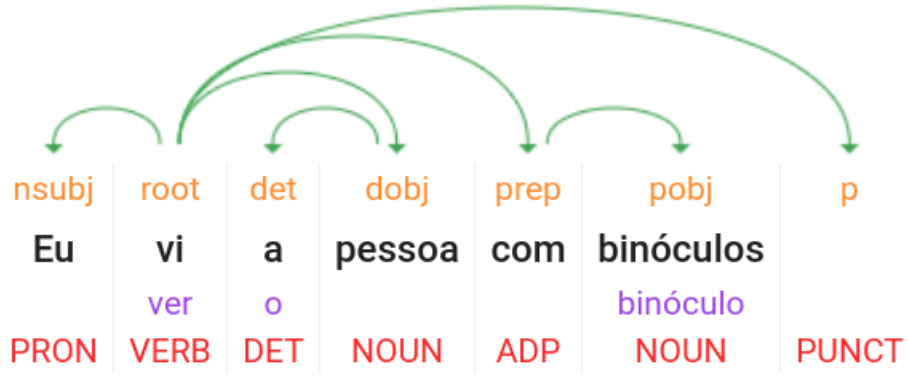
- Níveis (do mais simples para o mais complexo)
  - Morfológico
    - Tokenização, o que é uma palavra
  - Morfossintático
    - Flexões observadas diretamente na palavra
  - Sintático
    - Árvores sintáticas, substantivo, verbos, etc.
  - Semântico
    - Qual o significado da palavra ou frase
  - Discursivo
    - Qual a intenção do autor naquele texto/fala
  - Pragmático
    - Qual o contexto que está fora do texto/fala

- Há também aplicações de processamento de fala

#Níveis de Processamento

- Eu vi a pessoa com binóculos
  - Quem está com o binóculos?

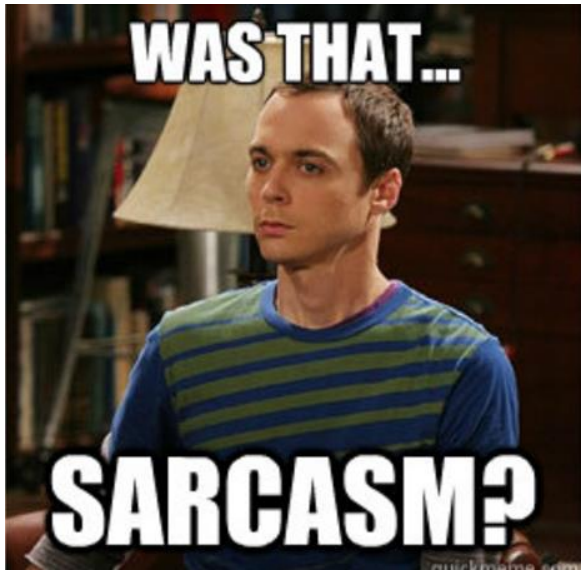
- Tokenização
  - [Eu, vi, a, pessoal, com, binóculos]
- Árvore sintática



- Semântica; Discurso e Pragmática?

#Níveis de Processamento – Problemas complexos

- Eu vi a pessoa com binóculos
  - Quem está com o binóculos?
- Sarcamos
- Joguei a xícara na mesa e ela quebrou?
  - O que foi quebrado?
  - Conhecimento de mundo
- O banco quebrou =/
  - Numa empresa de Tecnologia, talvez o banco de dados parou de funcionar
  - Mas pode ser um banco de madeira que estava em algum lugar...
  - Talvez, alguma instiuição financeira



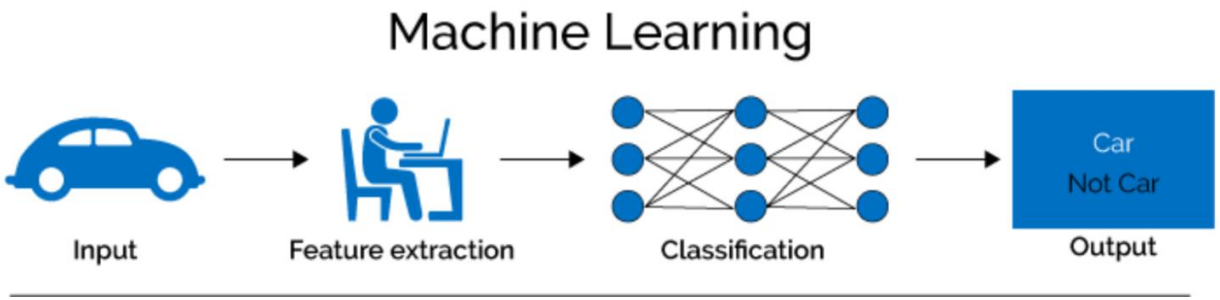
SiDi

PLN e Aprendizado de Máquina

Jul/2019

Aprendizado de Máquina

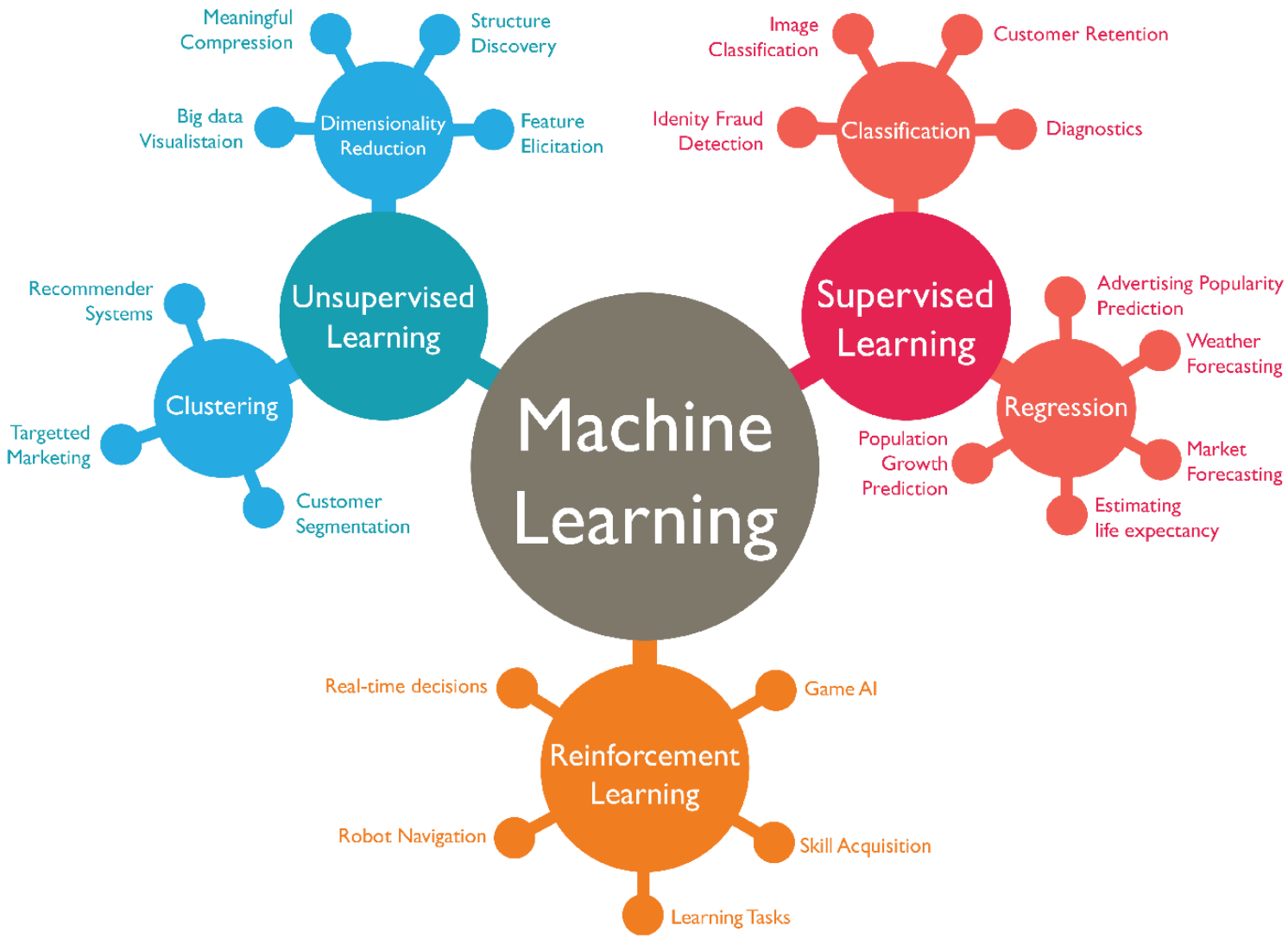
- Aprendizado de Máquina (AM) é uma área da IA que visa criar aplicações que **"aprendem" com experiência (dados) sem programação direta**
  - A máquina aprende um algoritmo sem ser programada



Por que usar Aprendizado de Máquina?

- Alguns problemas (as vezes "simples") exigem soluções muito complexas
  - Como fazer um computador reconhecer uma maça?

Aprendizado de Máquina



#Como fazer a máquina "entender" o texto?

- Transformar o texto em informação numérica

- Bag-of-Words
  - Countagem. TF-IDF, Binário

- Features manuais
  - A palavra inicia-se com letra maiúscula?
  - Quantas letras tem a palavra?
  - Quantas palavras tem a sentença?
  - Quantas sentenças tem o texto?
  - Etc.

- Representação vetorial (Deep Learning)

#Como fazer a máquina "entender" o texto?

- Input: Eu vi a pessoa com binóculos

- BOW:

outra	eu	vi	o	a	pessoa	professor	com	óculos	binóculos	...
0	1	1	0	1	1	0	1	0	1	...

- Features manuais:

# tokens	# verbos	# nouns	# uppercase	...
6	1	2	1	….

Token	Eu	vi	...
Upper case	Sim	Não	...
POS (classe gramatical)	Pronome	Verbo	...
# chars	2	2	..
Início?	Sim	Não	..

#Como fazer a máquina "entender" o texto?

- Cada tipo de problema requer um conjunto de features diferentes

- Corretor automático
  - Palavras anteriores
  - O que foi digitiado existe em algum dicionário?

- Chatbot
  - Localização da conversa
  - Contexto da conversa
  - Frase atual e anteriores
  - Etc.

#Deep Learning... uma simples (mesmo) apresentação

- Uso de redes neurais
- Features aprendidas automáticamente

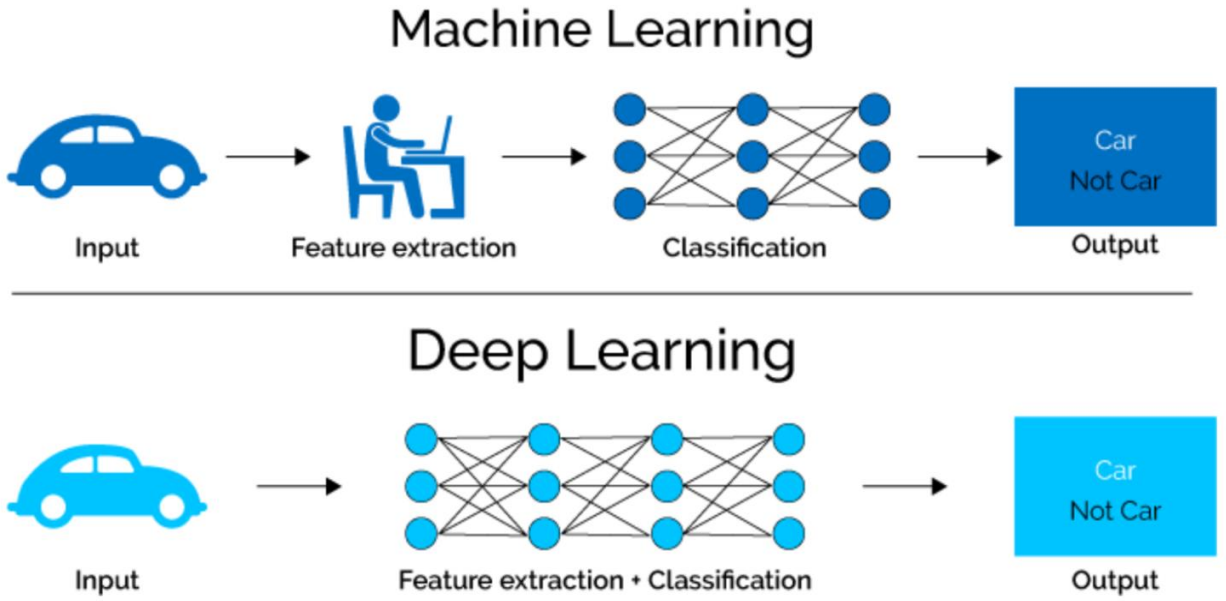


Figure 1: Machine Learning VS Deep Learning

#Deep Learning... como entender o texto

- Input: Eu vi a pessoa com binóculos

Eu	['0.13', '0.82', '0.56', '0.83', '0.41', '0.51', '0.07', '0.80', '0.98', '0.60']
vi	['0.57', '0.69', '0.41', '0.94', '0.93', '0.79', '0.12', '0.39', '0.65', '0.48']
a	['0.37', '0.25', '0.16', '0.10', '0.58', '0.19', '0.43', '0.27', '0.13', '0.89']
pessoa	['0.63', '0.48', '0.52', '0.53', '0.20', '0.79', '0.58', '0.17', '0.31', '0.98']
com	['0.39', '0.38', '0.75', '0.76', '0.48', '0.60', '0.99', '0.74', '0.75', '0.76']
binóculos	['0.61', '0.17', '0.21', '0.01', '0.39', '0.08', '0.40', '0.17', '0.37', '0.00']

- Palavras com significados similares apresentam vetores similares

SiDi

NTLK + Scikit-learn

Jul/2019

#NLTK

- Natural Language Toolkit (https://www.nltk.org/)
  - Diversas ferramentas para processamento de línguagem
    - Várias para português
  - Existem alguns módulos de Aprendizado de Máquina

- Scikit Learn (https://scikit-learn.org/stable/)
  - Biblioteca para Aprendizado de Máquina
  - Algoritmos, métodos de avaliação, etc;

- Pandas (https://pandas.pydata.org/)
  - Python Data Analysis Library

SiDi

Spayce

Jul/2019

#Spacy

- Industrial-Strength Natural Language Processing
  - Possui alguns modelos treinados para processamento
    - Inclusive para portuguese
- https://spacy.io/

#Informações gerais

- Github para o código de hoje:

- Alessandro Bokan: https://github.com/alessandrobokan
- Fernando Nóbrega: https://github.com/fernandoasevedo

https://www.sidi.org.br

Obrigado,

www.sidi.org.br