CURSO BÁSICO DE INTELIGÊNCIA ARTIFICIAL E BATE-PAPO COM CONVIDADOS ESPECIAIS

INTELIGÊNCIA ARTIFICIAL PARA TODOS

DE 08/06 A 12/08

COM OS PROFESSORES DO LABORATÓRIO ARIA/UFPB: TELMO FILHO, THAÍS GAUDENCIO E YURI MALHEIROS

Centro de Informática UFPB

Departamento de ESTATÍSTICA

aria artificial intelligence applications

- CURSO SEM PRÉ-REQUISITOS
- HTTP://ARIA.CI.UFPB.BR/IAPARATODOS
- INSCRIÇÃO PARA CERTIFICADO - DE 01/06 A 07/06: HTTP://BIT.LY/SIGEVENTOS
- ENCONTROS: SEGUNDAS E QUARTAS
- HORÁRIO: 19:00 ÀS 20:00

aria.ci.ufpb.br

SE INSCREVE E JÁ APERTA NO SININHO, QUE VOCÊS PASSAM A RECEBER AS NOTIFICAÇÕES.

NOSSOS ENCONTROS DURARÃO 1 HORA E, ASSIM QUE POSSÍVEL, DEIXAREMOS OS VÍDEOS GRAVADOS NO CANAL.

NÃO PRECISA SE PREOCUPAR EM ESTAR LIGADO ÀS 19:00, MAS ESTANDO, ROLA TIRAR DÚVIDA E PARTICIPAR, O QUE JÁ DEIXA A AULA MAIS ANIMADA.

SOBRE O MATERIAL DE ACOMPANHAMENTO: O ALUNO PRECISA SE LOGAR EM:
CLASSROOM.GOOGLE.COM
DEPOIS, CLICAR EM PARTICIPAR DA TURMA (ÍCONE COM UM MAIS +)

POR FIM, ENTRAR COM O CÓDIGO DA TURMA: PXV3ANW

TAMBÉM EM: HTTPS://ARIA.CI.UFPB.BR/IA-PARA-TODOS-MATERIAL/
ESPERAMOS QUE VOCÊS REALMENTE CURTAM O CURSO E APROVEITEM-NO AO MÁXIMO. NÃO DEIXEM DE INTERAGIR CONOSCO, TAMBÉM, POR E-MAIL OU MENSAGEM NO NOSSO INSTAGRAM (@APRENDIZAGEMDEMAQUINA)

Processamento de Linguagem Natural

Yuri Malheiros

Introdução

A linguagem é uma das mais importantes características humanas

A invenção da escrita é um marco importante na história da humanidade

Mudou a forma de armazenar e passar conhecimento

Facilita o ensino e aprendizagem

Introdução

A Internet tornou o conhecimento disponível em uma escala nunca vista antes

Hoje falta tempo para absorver tanto conhecimento

Introdução

Wikipédia:
+50 milhões de artigos

Amazon:
+6 milhões de ebooks

Wordpress:
~70 milhões de posts / mês

Introdução

Os seres humanos possuem excelente capacidade de compreensão do que é escrito, mas não temos velocidade de processamento

Os computadores possuem excelente velocidade de processamento, mas não têm boa capacidade de compreensão … por enquanto

Leitura

Uma das bases da leitura através de computadores é a contagem

O computador é excelente em processar números

É possível extrair conhecimento a partir de contagens de palavras

Leitura

Quais as palavras mais utilizadas por Machado de Assis em suas obras?

Retirando algumas palavras muito frequentes como artigos e preposições, temos:

É	4987	**Ainda**	1132
Disse	1430	**Ser**	1132
Casa	1236	**Nada**	1077
Olhos	1235	**Tempo**	1053
Tudo	1144	**Outra**	1026

Leitura

Também podemos contar pares (ou trios, …) de palavras:

Alguma coisa	341
Luís Garcia	215
Pode ser	212
Outra coisa	179
Podia ser	178
Quincas Borba	162

Vetorização

Os algoritmos de aprendizagem de máquina trabalham com entradas numéricas

Por isso, precisamos transformar textos em sequências de números (vetores)

A contagem é a base do modelo *Bag-of Words* (BoW)

Vetorização

O modelo BoW descreve a ocorrência das palavras em um texto

Textos similares devem ter representações similares no BoW

Dadas as frases abaixo, vamos criar a representação BoW delas:

1. Eu gostei do filme. O filme é divertido
2. Eu não assisti o filme

Vetorização

1. Eu gostei do filme. O filme é divertido
2. Eu não assisti o filme

Texto	eu	gostei	do	filme	o	é	divertido	não	assisti
1.	1	1	1	2	1	1	1	0	0
2.	1	0	0	1	1	0	0	1	1

Vetorização

Vamos interpretar cada texto como um vetor

“Eu gostei do filme. O filme é divertido” é representado pelo vetor:
(1, 1, 1, 2, 1, 1, 1, 0, 0)

“Eu não assisti o filme” é representado pelo vetor:
(1, 0, 0, 1, 1, 0, 0, 1, 1)

Vetorização

Note que como a ordem das palavras é perdida, frases diferentes podem ter a mesma representação no modelo BoW:

- O homem foi ao evento
- O evento foi ao homem

Ambas seriam representadas pelo mesmo vetor

Vetorização

Outro problema do BoW é que palavras que aparecem muito vão influenciar mais o resultado

Quais palavras são mais frequentes? Artigos, preposições …

Palavras que se repetem muito não costumam trazer muita informação

Vetorização

Para consertar isso, podemos penalizar palavras que aparecem muito em todos os textos analisados

Essa abordagem é chamada de *Term Frequency-Inverse Document Frequency* (TF-IDF)

Tokenização

Nos nossos exemplos, estamos trabalhando com palavras

Tokenização é o processo de quebrar um texto em pedaços menores chamados *tokens*

Tokenização

Muitas vezes os *tokens* correspondem as palavras de um texto

"Eu comprei pão” (tokens: eu, comprei, pão)

Mas nem sempre isso é claro

“Eu comprei R$ 10,00 de couve-flor” (quais os tokens?)

Tokenização

Em idiomas diferentes a tokenização pode ser mais difícil

Em inglês: “You aren’t old”

“aren’t" pode ser considerado um token, mas pode ser dividido em “are" “not"

Tokenização

Em idiomas diferentes a tokenização pode ser mais difícil

Em alemão, nomes compostos não são separados por espaços (ou hífens):

Lebensversicherungsgesellschaftsangestellter

Tradução: funcionário de uma empresa de seguro de vida

Classificação

Transformando texto em vetores, podemos usar algoritmos de aprendizagem de máquina

Por exemplo, classificação de texto de acordo com os seus sentimentos:

- Um dos melhores filmes de ação dos últimos anos 🙂
- Foi patético. A pior parte foram as cenas de luta 😕

Modelo de Linguagem

Indo além das contagens, podemos trabalhar com modelos probabilísticos

É possível prever quais as próximas palavras que alguém vai falar?

“Ao sair de casa desligue a _____ ”

Modelo de Linguagem

É possível prever quais as próximas palavras que alguém vai falar?

“Ao sair de casa desligue a _____ ”

É provável que a palavra seja “luz” ou “televisão”

Mas é bem menos provável que a palavra seja “geladeira” ou “bola”

Modelo de Linguagem

Qual frase é mais provável?

“Eu acendi o fósforo” ou “Eu ascendi o fósforo” ?

Modelo de Linguagem

Modelos que atribuem probabilidade a palavras ou sequências de palavras são chamadas de **modelos de linguagem**

Probabilidade de uma palavra: $P(w_1)$

Modelo de Linguagem

Modelos que atribuem probabilidade a palavras ou sequências de palavras são chamadas de **modelos de linguagem**

Probabilidade de uma palavra: $P(w_1)$

Probabilidade de uma sequência de palavras: $P(w_1, w_2, w_3, \ldots, w_n)$

Modelo de Linguagem

Modelos que atribuem probabilidade a palavras ou sequências de palavras são chamadas de **modelos de linguagem**

Probabilidade de uma palavra: $P(w_1)$

Probabilidade de uma sequência de palavras: $P(w_1, w_2, w_3, \ldots, w_n)$

Probabilidade da próxima palavra: $P(w_n|w_1, w_2, \ldots, w_{n-1})$

Modelo de Linguagem

As probabilidades são calculadas a partir do processamento de uma grande massa textual

As probabilidades carregam muito conhecimento

Modelo de Linguagem

Qual palavra está correta: “analisar" ou “analizar" ?

Podemos descobrir calculando a probabilidade de cada uma dessas palavras

P(analisar) > P(analizar)

Modelo de Linguagem

Qual frase está correta: “ele é maior" ou “ele é mais grande" ?

Podemos descobrir calculando a probabilidade de cada uma dessas frases

P(ele é maior) > P(ele é mais grande)

Modelo de Linguagem

Complete a frase: “A Terra é _____”

Modelo de Linguagem

Complete a frase: “A Terra é _____”

Probabilidades altas:

P(azul|A Terra é)

P(redonda| A Terra é)

Modelo de Linguagem

Complete a frase: “A Terra é _____”

Probabilidades altas:

P(azul|A Terra é)

P(redonda| A Terra é)

Probabilidades baixas:

P(amarela|A Terra é)

P(banana| A Terra é)

Modelo de Linguagem

Complete a frase: “A Terra é _____”

Probabilidades altas:

P(azul|A Terra é)

P(redonda| A Terra é)

Probabilidades baixas:

P(amarela|A Terra é)

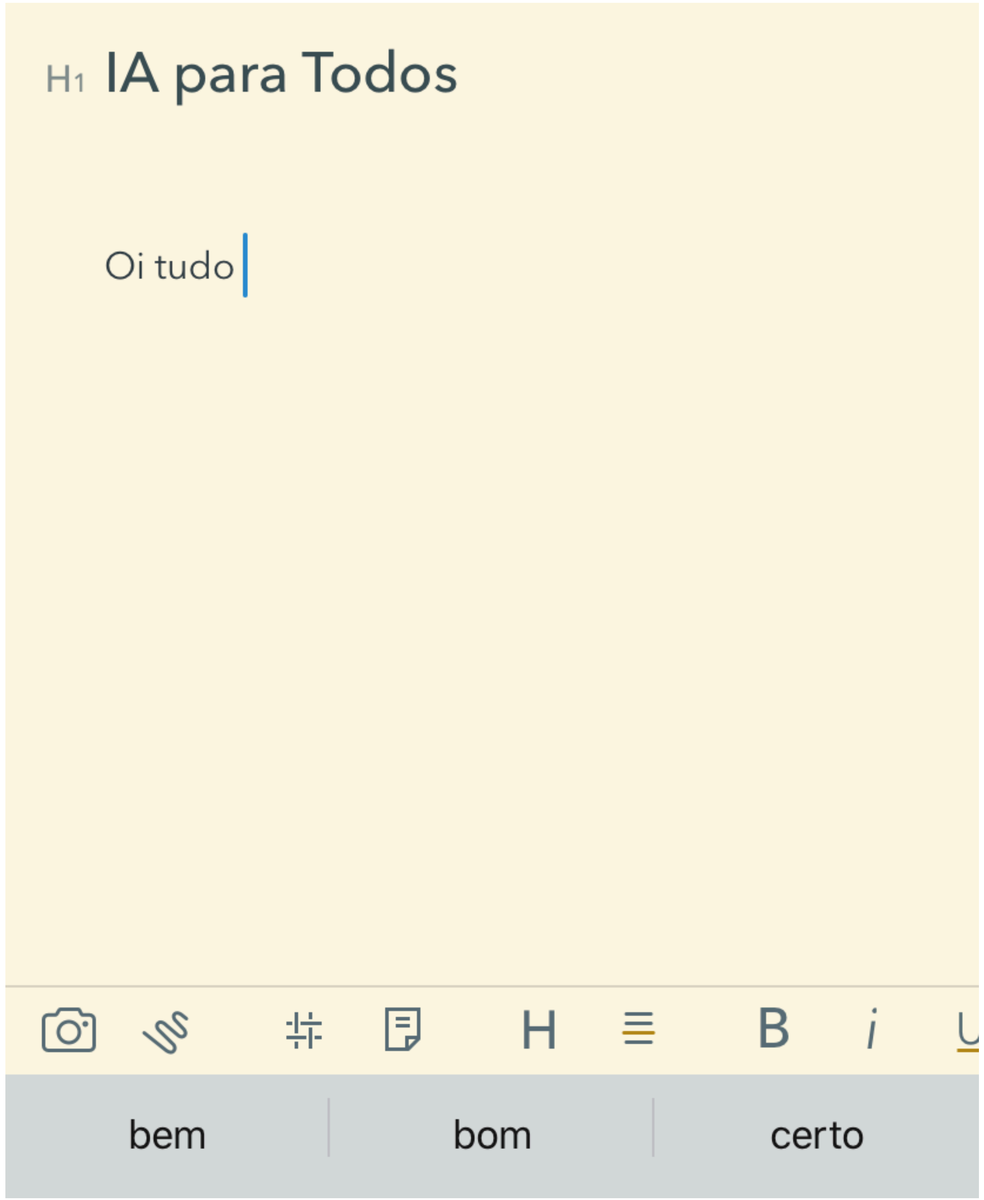
P(banana| A Terra é)

O modelo de linguagem incorpora conhecimentos gerais!

Modelo de Linguagem

Completar frases é a ideia por trás da sugestão de palavras dos celulares

Ele sugere as palavras que maximizam a probabilidade: P(w|Oi tudo)



Modelo de Linguagem

Computar as probabilidades é um desafio

Existem muitas possíveis combinações de palavras

Os modelos podem ficar muito complexos

Escrita

Podemos usar os modelos de linguagem para escrever automaticamente

Dado um texto inicial, qual palavra maximiza a probabilidade abaixo?

P(w|hoje tem) ?

Escrita

P(w|hoje tem) w = aula

Escrita

P(w|hoje tem) w = aula

P(w|hoje tem aula) w = de

Escrita

P(w|hoje tem) w = aula

P(w|hoje tem aula) w = de

P(w|hoje tem aula de) w = IA

Escrita

P(w|hoje tem) w = aula

P(w|hoje tem aula) w = de

P(w|hoje tem aula de) w = IA

P(w|hoje tem aula do IA) w = para

Escrita

P(w|hoje tem) w = aula

P(w|hoje tem aula) w = de

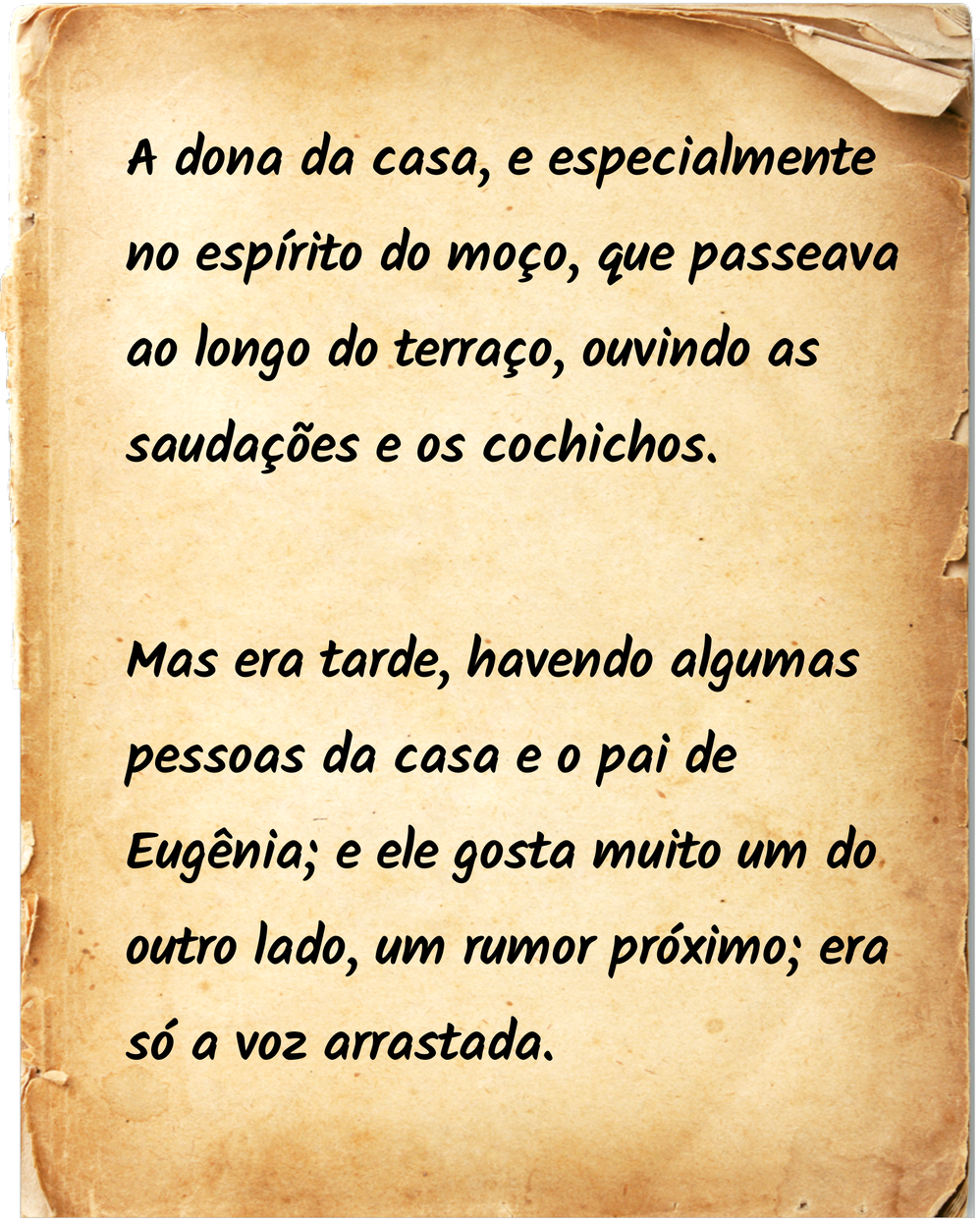
P(w|hoje tem aula de) w = IA

P(w|hoje tem aula do IA) w = para

P(w|hoje tem aula do IA para) w = Todos

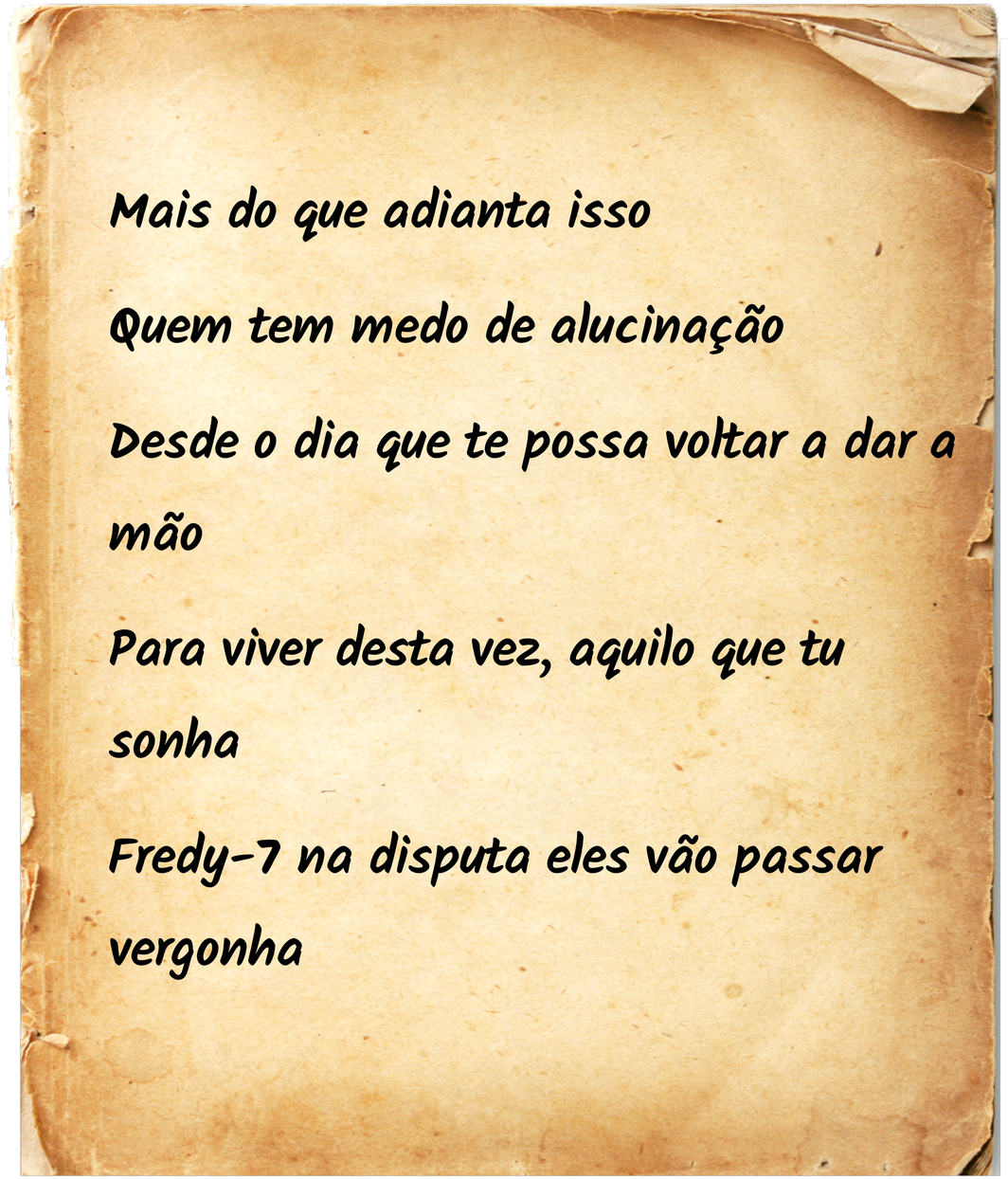
Escrita

Aplicando esta ideia para os textos de Machado de Assis:



Escrita

Aplicando a letras de Rap:



Escrita

Um dos modelos de linguagem mais avançados atualmente é o GPT-3

Foi criado usando mais de 1 trilhão de *tokens*

São aproximadamente 700 Gb para carregar o modelo na memória!

Escrita

Entrada: “Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.”

Saída: The orcs’ response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. “You are in good hands, dwarf,” said Gimli, who had been among the first to charge at the orcs;

Escrita

Entrada: “For today’s homework assignment, please describe the reasons for the US Civil War.”

Saída: By Donny Ferguson

It is easy to identify why the Civil War happened, because so many people and so many books and so much television and films tell us that it was the cause, that it has something to do with race or economics or religion…

Conclusão

Existem muitas outras áreas relacionadas ao processamento de linguagem natural

- Tradução
- Sumarização
- Responder perguntas
- Extração de conhecimento
- Resolução de correferência
- Reconhecimento de entidades nomeadas

Conclusão

Aplicações usando PLN estão cada vez mais presentes no dia a dia

Mas muito ainda precisa evoluir

Representação vetorial de textos e modelos de linguagem ainda continuam melhorando

@aprendizagemdemaquina

BATE-PAPO COM HERMANO VIANNA

05/08 - 19:00 - AO VIVO

YOUTUBE.COM/ARIALAB

CULTURA E IA

Hermano Vianna é antropólogo, pesquisador musical e roteirista de televisão. Autor dos livros "O Mistério do Samba" (Zahar, 1995) e "O Mundo Funk Carioca" (Zahar, 1988) é também criador dos programas Esquenta!, Central da Periferia, Brasil Legal e Programa Legal (TV Globo) e foi colunista do Jornal O Globo. Em 2005 foi criador do Overmundo, um website colaborativo com o objetivo de dar visibilidade na internet à produção cultural que não é vista na grande mídia e contou com mais de 1 milhão de visitantes únicos por mês. Curioso e leitor assíduo sobre tecnologia, escreve no https://hermanovianna.wordpress.com/