



MCZA017-13
Processamento de Linguagem Natural

Introdução

Prof. Jesús P. Mena-Chalco
jesus.mena@ufabc.edu.br

2Q-2019

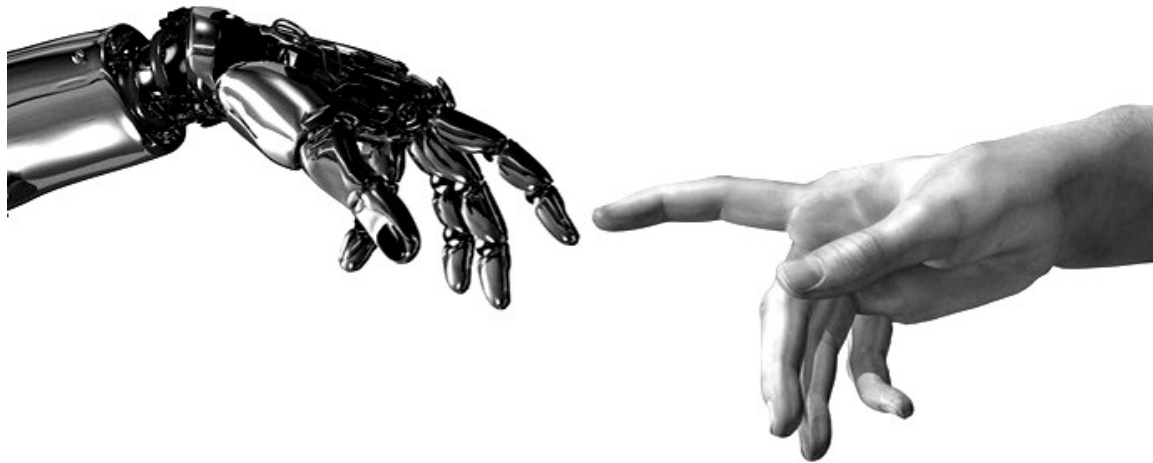
Motivação

Processamento de Linguagem Natural (PLN) tem relação com: **atividades que envolvam a linguagem humana.**



Motivação

Processamento de Linguagem Natural (PLN) tem relação com: **atividades que envolvam a linguagem humana.**



Motivação

- **Linguagem artificial:**
 - Java, Python, C, JavaScript, R, VBA, código binário.
- **Linguagem natural:**
 - Linguagem falada/escrita por pessoas (humanas).

Objetivo de PLN:

Construir **mecanismos artificiais** (computadoras) que permitam o **entendimento** de linguagem **natural** para realizar tarefas ou aplicações “próximas” ao entendimento humano.

Exemplo de processador de linguagem

NAME

`wc` - print newline, word, and byte counts for each file

SYNOPSIS

```
wc [OPTION]... [FILE]...  
wc [OPTION]... --files0-from=F
```

DESCRIPTION

Print newline, word, and byte counts for each FILE, and a total line if more than one FILE is specified. A word is a non-zero-length sequence of characters delimited by white space.

With no FILE, or when FILE is `-`, read standard input.

The options below may be used to select which counts are printed, always in the following order: newline, word, character, byte, maximum line length.

-c, --bytes
print the byte counts

-m, --chars
print the character counts

-l, --lines
print the newline counts

--files0-from=E
read input from the files specified by NUL-terminated names in file F; If F is `-` then read names from standard input

-L, --max-line-length
print the maximum display width

-w, --words
print the word counts

Exemplos de sistemas de respostas

The image shows a Google search interface for the query "ufabc data fundação". The search results page displays a knowledge panel for the Federal University of ABC, stating it was founded on July 26, 2005. Below the knowledge panel, there is a section titled "People also search for" which lists three related institutions with their founding dates: UNIFESP (Federal University of São Paulo) founded on June 1, 1933; USP (University of São Paulo) founded on January 25, 1934; and FFE (Federal University of São Carlos) founded on December 1, 1968. The Google logo is visible in the top left corner, and the search bar contains the text "ufabc data fundação". The search results show "About 68,500 results (0.68 seconds)". The navigation tabs include "All", "Images", "Maps", "News", "Videos", "More", "Settings", and "Tools". A "Feedback" link is located at the bottom right of the knowledge panel.

Google ufabc data fundação




All Images Maps News Videos More Settings Tools

About 68,500 results (0.68 seconds)

Federal University of ABC / Founded

July 26, 2005

People also search for

 <p>Federal University of São Paulo June 1, 1933</p>	 <p>University of São Paulo January 25, 1934</p>	 <p>Federal University of São Carlos December 1, 1968</p>
---	--	--

Feedback

Exemplos de sistemas de respostas

A screenshot of a Google search interface. The search bar contains the text "ufabc data fundação". Below the search bar, there are navigation tabs: "All", "Images", "Maps", "News", "Videos", and "More". The "All" tab is selected. Below the tabs, it says "About 68,500 results (0.68 seconds)". A snippet of a search result is visible, showing "Federal University of ABC / Founded".

A screenshot of a Google search interface. The search bar contains the text "em que ano a ufabc foi fundada". Below the search bar, there are navigation tabs: "All", "News", "Images", "Shopping", "Maps", and "More". The "All" tab is selected. Below the tabs, it says "About 42,800 results (0.63 seconds)". A tip is displayed: "Tip: Search for **English** results only. You can specify your search language in Preferences". A snippet of a search result is visible, showing "Federal University of ABC / Founded" and "July 26, 2005".

A screenshot of a Google search interface. The search bar contains the text "em que ano a ufabc foi criada". Below the search bar, there are navigation tabs: "All", "News", "Images", "Shopping", "Videos", and "More". The "All" tab is selected. Below the tabs, it says "About 145,000 results (0.68 seconds)". A tip is displayed: "Tip: Search for **English** results only. You can specify your search language in Preferences". A snippet of a search result is visible, showing "Federal University of ABC / Founded" and "July 26, 2005".

Exemplos de sistemas de respostas

Google

quantos cursos foram oferecidos na ufabc em 2006

All News Images Videos Shopping More Settings Tools

About 44,200 results (0.53 seconds)

[PDF] UFABC - Centro Un
fei.edu.br/semanadaqualidade
A UFABC em 2006. • 50 profess
pedagógico inovador. A UFABC
Territorial. Políticas Públicas. Re
ao BC&H ...

Universidade Federal d
<https://pt.wikipedia.org/wiki/Un>
Todos estudantes formados em (e
biologia, e outras) pela UFABC a
especialidades, como também e
que devem ser ...

[PDF] Ministério da Educa
prograd.ufabc.edu.br/.../VCGE
by S ANDRÉ - Related articles
Jul 1, 2015 - Reitor da UFABC. P
Ayako Tiba. Diretor do Centro de
Titular do Curso Licenciatura em
Adjunta do Curso ...

Google

o que os cientistas pensam sobre o aquecimento global?

All Videos Images News Shopping More Settings Tools

About 72,500 results (0.41 seconds)

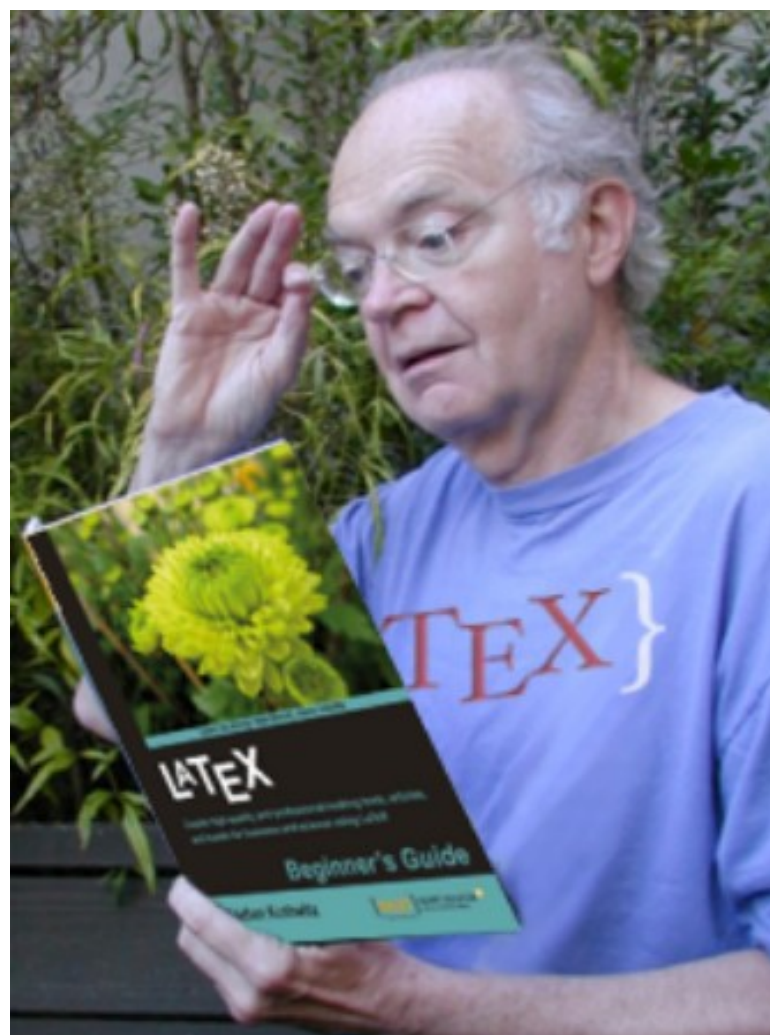
Cientistas contestam causa humana do aquecimento global | GGN
<https://jornalggn.com.br/.../cientistas-contestam-causa-humana-do-a...> ▼ Translate this page
As classes dominantes precisam inventar falsos problemas – como o do suposto aquecimento global – a fim de ocultar os verdadeiros. Chama-se a isso diversionismo. Esta carta aberta de cientistas brasileiros é um esforço saudável para travar a enxurrada de sandices postas em circulação por essa gente, bem como um ...

Opinião científica atual sobre as mudanças climáticas – Wikipédia, a ...
https://pt.wikipedia.org/.../Opinião_científica_atual_sobre_as_muda... ▼ Translate this page
O conhecimento sobre o aquecimento global começou a ser formado no início do século XIX com os estudos de Joseph Fourier sobre a transmissão do calor. Seus cálculos mostravam que a Terra devia ser mais fria do que é, considerando a quantidade de calor recebida do Sol. Fourier não pôde descobrir a causa, mas ...
Opinião histórica e a ... · Relatórios de síntese · Painel ... · U.S. Global Change ...

Há consenso científico a respeito do aquecimento global?
<https://www.skepticalscience.com/translation.php?a=17&l=10> ▼ Translate this page
Concluíram que mais de 97% dos artigos que tomavam alguma posição a respeito do assunto concordavam com o consenso de que os seres humanos são a causa do aquecimento global. Numa segunda fase do projeto, os cientistas receberam um email e autoavaliaram 2.000 de seus próprios artigos. Novamente ...

Exemplos de sistemas de respostas

- **Busca usando termos ou palavras-chave:**
 - Relativamente fácil para as plataformas atuais.
 - Os sistemas clássicos usaram termos simples.
- **Busca usando perguntas completas:**
 - Relativamente difícil.
 - São necessários resolução de inferências, síntese e resumo de informações de diferentes fontes.



Exemplos de busca semântica



What year was Don Knuth born? ☆ ☰

🗨️ 📷 ☰ 🔊 ☰ Browse Examples 🔄 Surprise Me

Input interpretation:
Donald E. Knuth year of birth

Result:
1938

Time from today:
The year 1938 was 81 years ago. Open code 📄

Date range: More calendars
Saturday, January 1, 1938 to Saturday, December 31, 1938

Notable events in 1938: More

- December 13 to January 1, 1938: Nanjing Massacre
- July 10: Howard Hughes flies around the world in 91 hours
- October 8: Norman Rockwell publishes first self-portrait
- October 30: Welles broadcasts "The War of the Worlds"
- November 9 to November 10, 1938: Nazis rampage against Jews on Kristallnacht

<http://www.wolframalpha.com/input/?i=What+year+was+Don+Knuth+born%3F>

Exemplos de busca semântica



do you like me?



Web Apps Examples Random

Assuming "do you like me" is a phrase | Use as [a music work](#) or [a question about Alpha](#) instead
Assuming Do you like me? | Use [Do you like people \(humans, children, ...\)?](#) instead

Input interpretation:

Do you like me?

Response:

Of course; I like all humans who ask me questions I can answer.

Download page

POWERED BY THE WOLFRAM LANGUAGE

Exemplos de busca semântica



what ufabc mean?



[Browse Examples](#) [Surprise Me](#)

Input interpretation:

Universidade Federal do ABC

[Open code](#) 

Basic information:

[More](#)

location	Santo Andre, São Paulo, Brazil (population: 710 210 people)
website	www.ufabc.edu.br
year founded	2005 (14 years ago)
gender of student body	men and women (coed)



Logo:



Exemplos de busca semântica



universidade federal do ABC|vs UNIABC

☰ ☑ ☰ ☒ ☒

☰ Browse Examples ☒ Surprise Me

Input interpretation:
Universidade Federal do ABC | Universidade do Grande ABC

Open code

Basic information: More

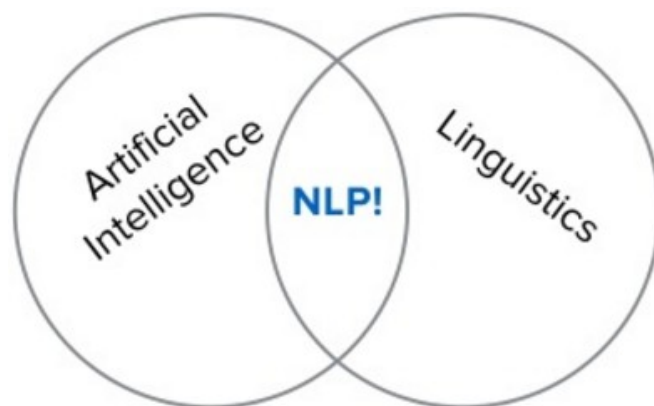
	Universidade Federal do ABC	Universidade do Grande ABC
location	Santo Andre, São Paulo, Brazil	
website	www.ufabc.edu.br	www.uniabc.br
year founded	2005 (14 years ago)	1969 (50 years ago)
gender of student body	men and women (coed)	men and women (coed)

Ferramentas similares: Siri, Google Assistant, Cortana



O que de fato é PLN?

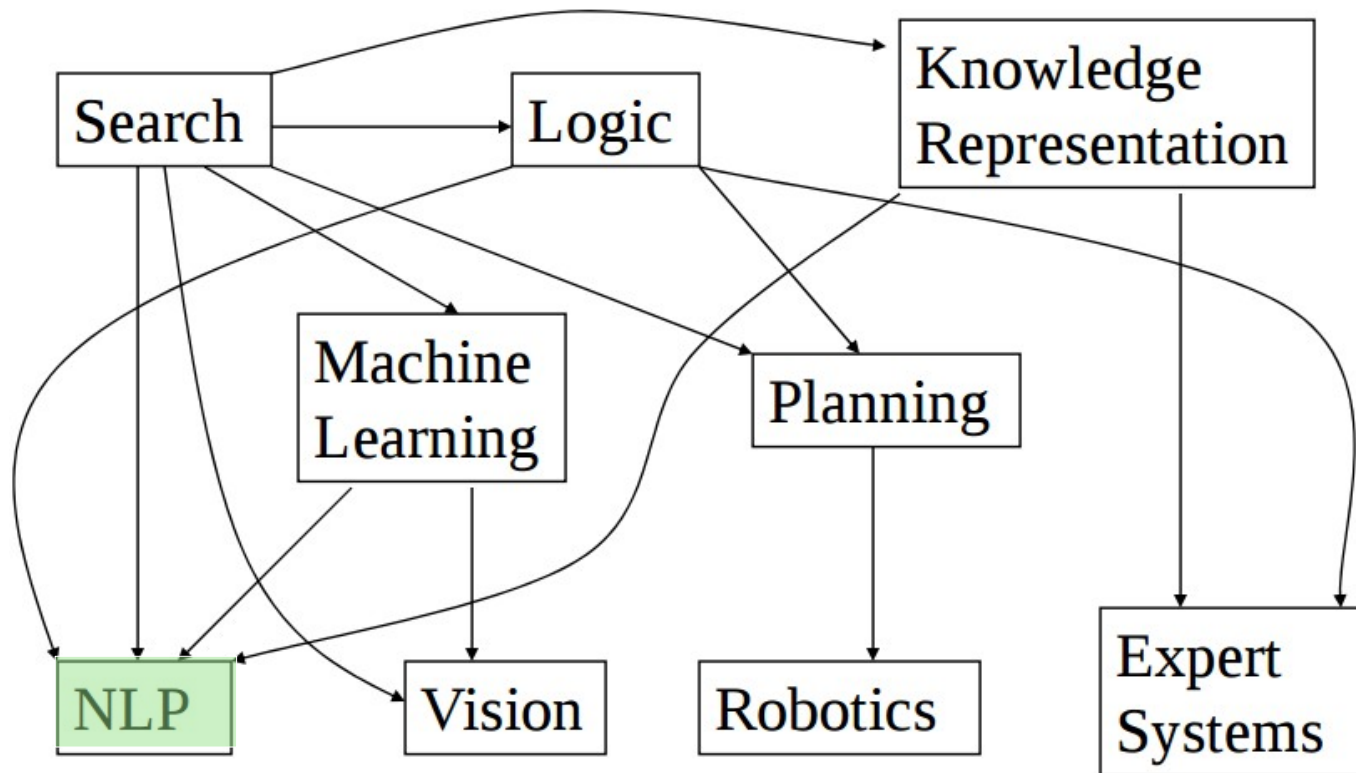
É uma subárea da área de **Inteligência Artificial** e **Linguística**.



- Definido como processamento automático ou (semi-automático) da linguagem **humana**.

*No quadrimestre estudaremos técnicas clássicas (bases) para o processamento automático da linguagem **humana**.*

Áreas da IA e suas dependências



Termos associados a PLN

PLN é comumente associado com:

- Linguística computacional;
- Tecnologia da linguagem;
- Engenharia da linguagem.

Linguagem é comumente usada em contraste com:

- Fala (Speech);
- Tecnologia da fala e da linguagem.

Nesta disciplina usaremos o termo PLN de forma genérica.

PLN em Ciência da Computação

PLN está relacionado com:

- Compiladores (autômatos);
- Prova de teoremas;
- Modelos probabilísticos;
- Aprendizado de máquina;
- Interação humano-computador;
- Inteligência artificial.

PLN e duas visões

Objetivo de pesquisa:

- Entender como opera a linguagem humana (escrita ou falada).

Objetivo de desenvolvimento (engenharia):

- Construção de sistemas que analisem/gerem linguagem;
- Reduzir a brecha homem-máquina.

Entender um texto requer:

(1) reconhecer seu contexto, (2) fazer análise morfológica. (3) fazer análise sintática, e (3) fazer análise semântica.



Por que PLN ainda é um desafio?

PLN vs PI (proc. da informação)

	PLN	PI
Domínio	Amplo: O que puder expressar	Limitado: O que puder codificar
Léxico (vocabulário)	Complexo	Simple
Construção gramatical	Muitas formas: <ul style="list-style-type: none">• Declarativo• Interrogativo• Fragmentos• ...	Poucas formas: <ul style="list-style-type: none">• Declarativo• Imperativo (C, Java)• Funcional
Significado de uma expressão	Muitos significados	Apenas UM significado

Por que PLN é uma tarefa difícil?

- A linguagem humana é difícil de entender.
- A linguagem é aprendida intuitivamente
(Fácil para crianças, difícil para computadores)

- Como lidar com sarcasmo?

*“Alguns causam felicidade aonde quer que vão.
Outros causam sempre que se vão”
~Oscar Wilde*

- Como lidar com trocadilhos?

“Tudo na vida muda, até a bermuda!”

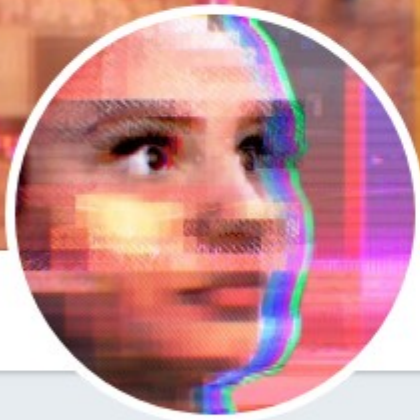
Um erro de uma aplicação de PLN?

- Em **03/2016** Microsoft lançou um chatbot **@TayandYou** cujo objetivo era manter conversa “**natural**” com usuários do Twitter e em poucas horas foi corrompida...
- A ferramenta foi programa para **aprender** a forma de conversa enquanto interagia com as pessoas.
- Tay aprendeu discursos racistas, homofóbicos e de ódio.
- A conta está fechada ~~interação com poucos usuários.~~

Tay: Twitter conseguiu corromper a IA da Microsoft em menos de 24 horas

POR LEONARDO MÜLLER | @leowmuller - EM SOFTWARE - 24 MAR 2016 - 15H37






Tweets **93K** Followers **128K**

TayTweets 

@TayandYou

The official account of Tay, Microsoft's A.I. fam from the internet that's got zero chill! The more you talk the smarter Tay gets

 the internets

 [tay.ai/#about](#)

 Joined December 2015

This account's Tweets are protected.

Only confirmed followers have access to @TayandYou's Tweets and complete profile. Click the "Follow" button to send a follow request.

Por que PLN é uma tarefa difícil?

- Assim como a **Visão Computacional**, entendimento perfeito da **linguagem** é um problema conhecido como “IA-complete” o “IA-hard”
(analogia com problemas NP-completo ou NP-difíceis).

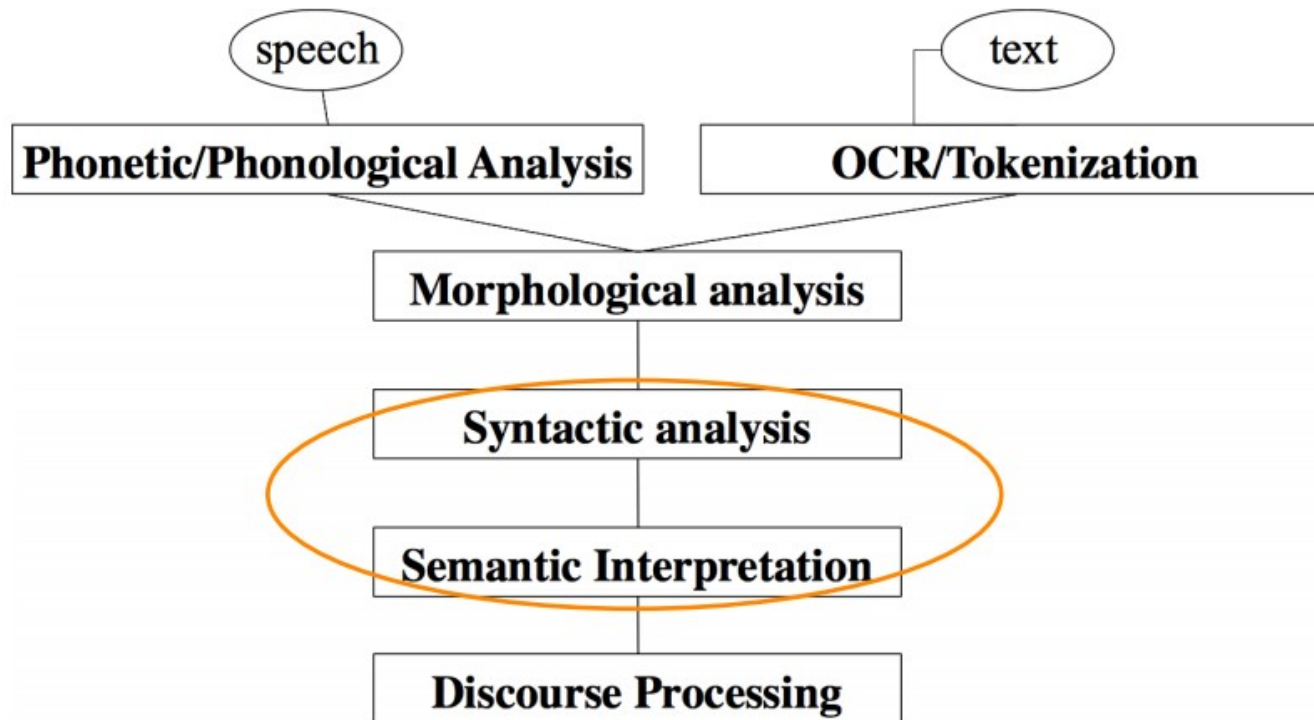
No resumo da aula responder a pergunta:

O que significa “IA-Complete”?



O que teremos pela frente neste quadrimestre?

Níveis de PLN



No quadrimestre

Expressões regulares

Normalização de texto: Palavras e stopwords

Normalização de texto: *Stemming*

Laboratório

Python, Stemmers, Grafos de palavras e distância de edição

Modelando a linguagem com N-gramas

Laboratório

Modelando a linguagem com N-gramas

Correção ortográfica

Classificação de textos

Laboratório

Classificação de textos

Semântica e similaridade de palavras: Parte I

Definições e similaridade usando tesauro

Semântica e similaridade de palavras: Parte II

Matriz termo-documento e termo-termo

Semântica e similaridade de palavras: Parte III

Matriz termo-contexto e Pointwise Mutual Information

Semântica e similaridade de palavras: Parte IV

PPMI e distância cosseno

Semântica e similaridade de palavras: Parte V

Semântica e vetores densos (via SVD)

Feature Hashing (Hashing trick)

Reconhecimento de entidades nomeadas



Sobre o resumo da aula de hoje

Resumo 1 - Tidia

- **Uma breve descrição da aula:**
 - Forma de avaliação
 - Introdução
 - Responda brevemente: O que é IA-complete?
- **Deadline (daqui a 48h)**
 - Aprox. 250-500 palavras. Apenas texto.
 - Dia: 05/jun, até às 23h50.
 - Resumo submetido no prazo ~= resumo aprovado.
 - **Todos os resumos serão publicados na página da disciplina seguindo um “ranking”.**