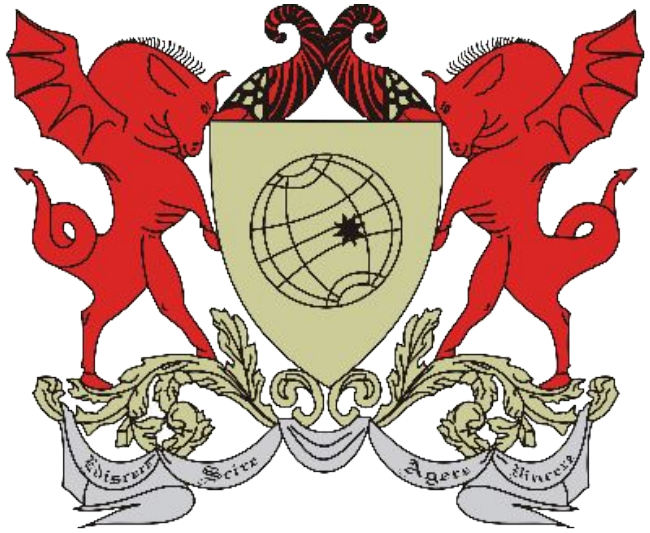


LINGUÍSTICA COMPUTACIONAL

Alcione de Paiva Oliveira - DPI/UFV

LINGUÍSTICA COMPUTACIONAL

Sumário

- Introdução
- Dificuldades para PLN
- Breve História
- Por que pesquisar em LC
  - Base teórica
  - Pesquisas na UFV
  - Ferramentas

http://blog.wordbank.com/2013/08/28

LINGUÍSTICA COMPUTACIONAL

Introdução

LINGUÍSTICA COMPUTACIONAL

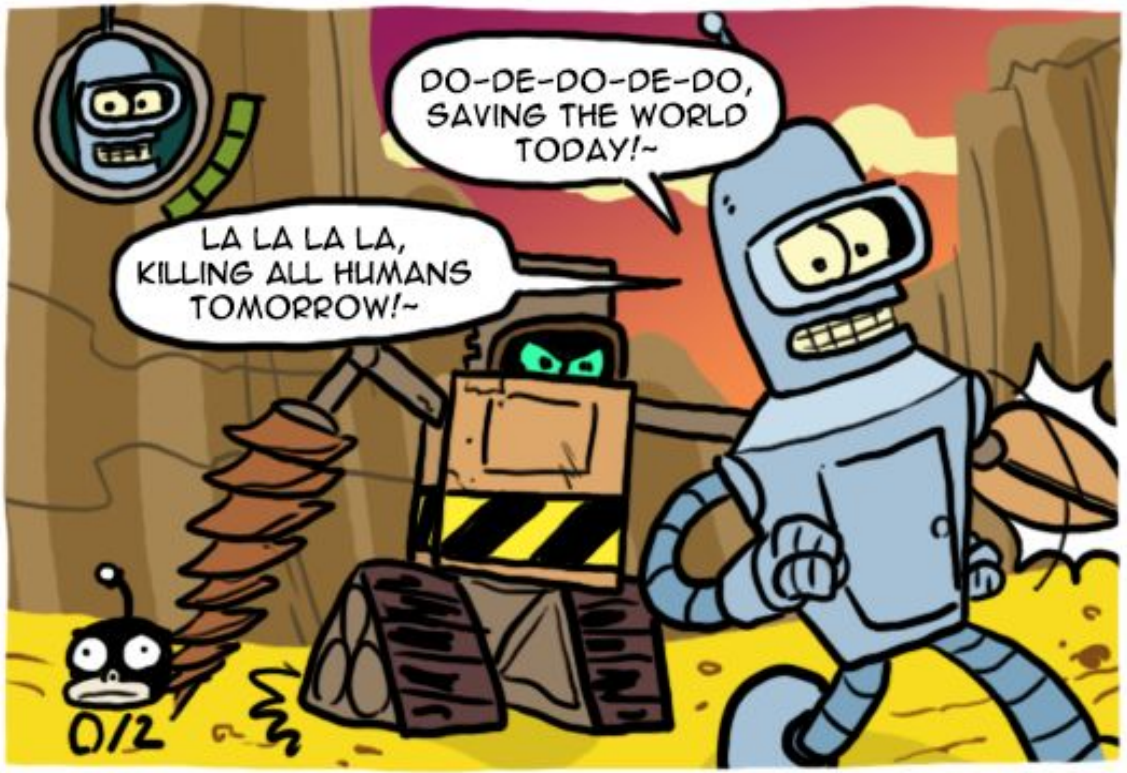
Esta palestra trata de uma área interdisciplinar.

Possui várias denominações:

- **linguística computacional**
- **Processamento da Linguagem Natural (PLN)**
- **Tecnologia da Linguagem humana.**

LINGUÍSTICA COMPUTACIONAL

O objetivo deste campo de pesquisa é conseguir que os computadores sejam capazes de fazer algo útil usando a linguagem humana.



LINGUÍSTICA COMPUTACIONAL

Não é uma especialização da Linguística.

O objeto de estudo da linguística é a linguagem natural em seu vários aspectos (surgimento, uso, ligação com a cognição, aspectos sociais, etc...)

Estudos atuais mostram que a linguagem e o pensamento possuem uma ligação.

LINGUÍSTICA COMPUTACIONAL

Linguagem e pensamento.

LINGUÍSTICA COMPUTACIONAL

Linguagem e pensamento.

Que adjetivos você usaria para uma ponte?

LINGUÍSTICA COMPUTACIONAL

Áreas relacionadas

- **Reconhecimento de voz** - trata de reconhecer os sons, gerando como saída as entidades léxicas.
- **Sintetização de voz** - transforma entidades léxicas e frases e sons.
- **Compreensão da linguagem natural** - trata de entender o significado de sentenças em linguagem natural.
- **Geração de textos** - trata da geração de sentenças ininteligíveis por pessoas que transmitam o significado associado a uma determinada representação.
- **Tradução automática** – tradução entre linguagens naturais.
- **Categorização de textos**
- **Análise de sentimento**

Dificuldades para PLN

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- As frases de uma língua são descrições incompletas das informações que pretendem transmitir

Há alguns cachorros lá fora

- Há alguns cachorros no jardim.
- Há três cachorros no jardim.
- Nick, Dingo e Sheik estão no jardim.

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- As frases de uma língua são descrições incompletas das informações que pretendem transmitir

Há alguns cachorros lá fora

- Há alguns cachorros no jardim.
- Há três cachorros no jardim.
- Nick, Dingo e Sheik estão no jardim.

A mesma expressão significa coisas diferentes em contextos diferentes:

Vou à praia frequentemente	Se for carioca significa no mínimo toda semana
Vou à praia frequentemente	Se for mineiro significa no máximo uma vez mês.

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- A mesma expressão significa coisas diferentes em contextos diferentes:

O servidor caiu

- Sala de monitores

- 

- Esta casa é uma zona

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- Nenhum programa de LN pode ser completo porque novas palavras, expressões e significados podem ser gerados com bastante liberdade:

Vou deletar você.

- Há inúmeras maneiras de dizer a mesma coisa:

Maria nasceu no dia 11 de outubro.
O aniversário de Maria é no dia 11 de outubro.

- Uma sentença pode ser interpretada de diversas formas:

Eu vi Maria fumando de binóculos
Isso aconteceu quando Maria estava grávida do Lucas

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- As sentenças sobre um evento variam em função da perspectiva:

Vou usar remédio para barata.

Perspectiva da pessoa.

Vou usar veneno para barata.

Perspectiva da barata.

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- A ordem dos elementos muda o significado:

A bota a gente calça

A calça a gente bota

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- Existem muitas figuras de linguagem
  - **Metonímia**
  - Enunciado: *João comeu um prato de comida.*
  - Significado: *João comeu a comida que estava no prato.*

  - **Metáfora**
  - Enunciado: *aquele médico é um açougueiro.*
  - Significado: *aquele medico operou mal.*

LINGUÍSTICA COMPUTACIONAL

Problemas do PLN

- Palavras sem significado e marcadores de discursos
  - **Marca**
  - Enunciado: *Veja bem, isto é uma variável, né?*
  - Significado: *isto é uma variável.*
  -

LINGUÍSTICA COMPUTACIONAL

Como é compreendido o significado de um lexema?

Substituir

*Robinho **substituiu** Neymar*

<jogador> *<jogador>*

LINGUÍSTICA COMPUTACIONAL

Como é compreendido o significado de um lexema?

Substituir

*Mano Menezes **substituiu** Neymar*

<Técnico> *<jogador>*

Baseado em exemplo de Margarida Salomão

LINGUÍSTICA COMPUTACIONAL

Quando podemos dizer que uma sentença foi compreendida?

1) Quando o ouvinte cria um modelo mental que corresponde ao modelo que o falante desejava transmitir.

2) Quando o o ouvinte realiza a ação que o falante solicitou em seu enunciado.

LINGUÍSTICA COMPUTACIONAL

Breve História

LINGUÍSTICA COMPUTACIONAL

No final dos anos 1950 e início dos anos 1960 o estudo processamento da linguagem natural se separou claramente em dois paradigmas: simbólicos e estocásticos.

LINGUÍSTICA COMPUTACIONAL

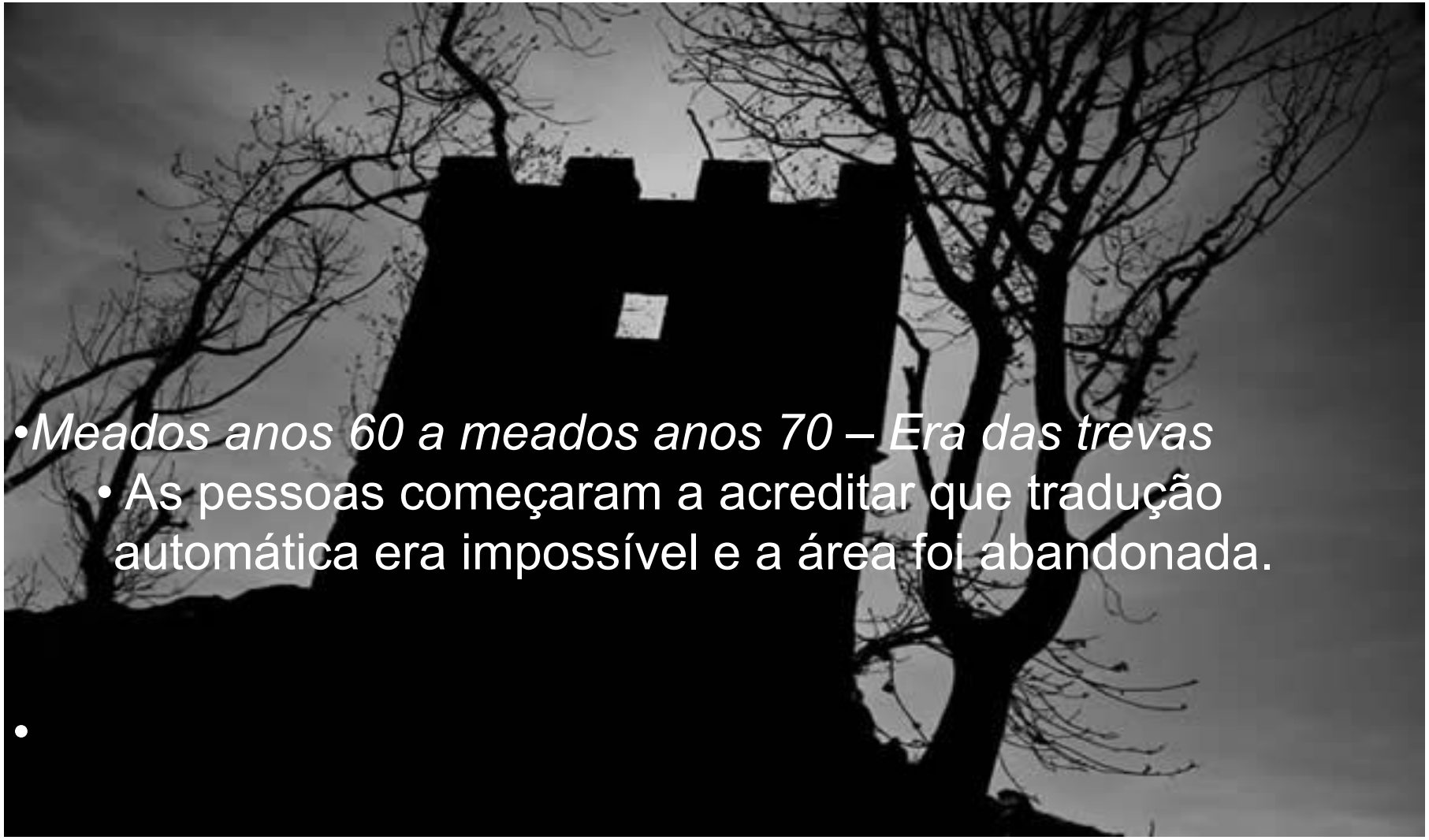
Breve História

- *Meados anos 50 a meados anos 60*
  - Nascimento do PLN – inicialmente os pesquisadores acreditavam que PLN era fácil e previram que a tradução por máquina seria resolvida em 3 anos.
  - Regras codificadas à mão e abordagens baseadas em teorias linguísticas (Gerativismo - Chomsky)

```
<sent> → <SN><SV>
<SN>   → <DET><SUB>
<SV>   → <V><S>
```

LINGUÍSTICA COMPUTACIONAL

Breve História



- *Meados anos 60 a meados anos 70 – Era das trevas*
  - As pessoas começaram a acreditar que tradução automática era impossível e a área foi abandonada.
-

LINGUÍSTICA COMPUTACIONAL

Breve História

- *Anos 70 até início dos anos 80* - Retomada
  - As atividades de pesquisa são reiniciadas, mas ainda com base nas teorias linguísticas gerativas e com pequenos problemas.

LINGUÍSTICA COMPUTACIONAL

Breve História

- *Anos 90 – Revolução estatística*
  - O poder computacional cresceu enormemente.
  - A abordagem estatística dirigida por dados com representação simples se mostra superior que o uso da abordagem baseada em regras complexas.
  - Existe uma certa semelhança com as novas teorias linguísticas cognitivas (Fillmore e Lakoff)

Sempre que demito um linguísta o desempenho da nossa máquina de tradução melhora. (Jelinek, 1988)

LINGUÍSTICA COMPUTACIONAL

Breve História

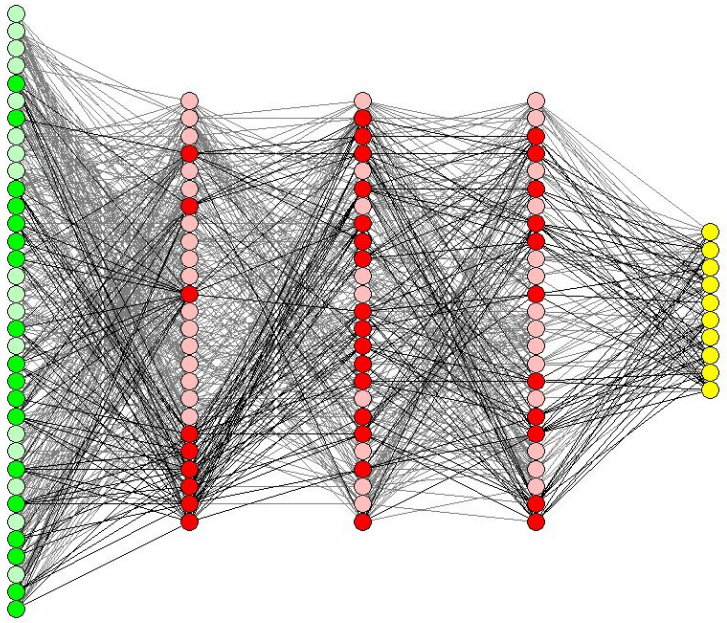
- *Anos 2000 – Estatística com contribuições da linguística*
  - Representação de dados e modelos estatísticos mais sofisticados

LINGUÍSTICA COMPUTACIONAL

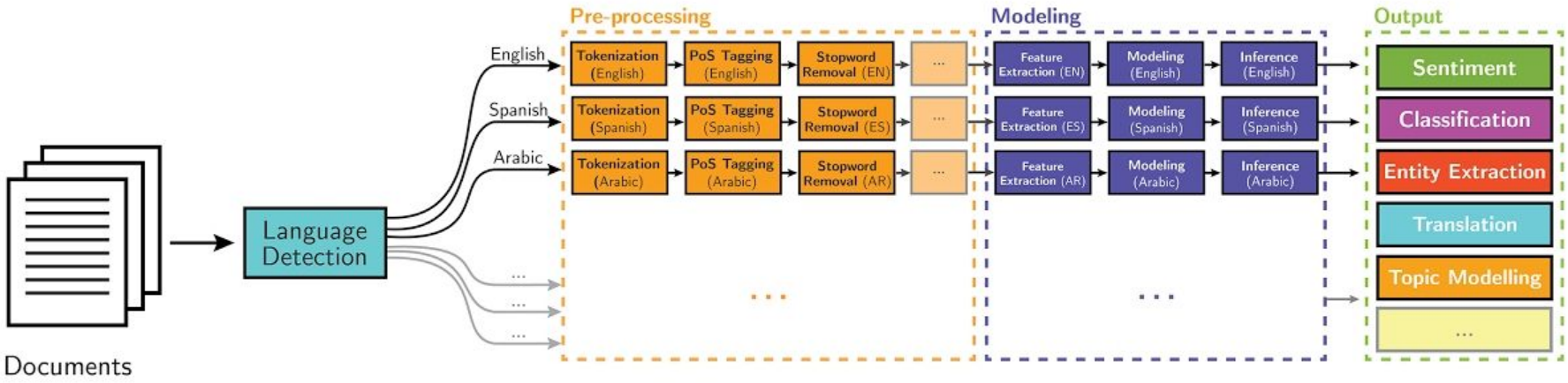
Breve História

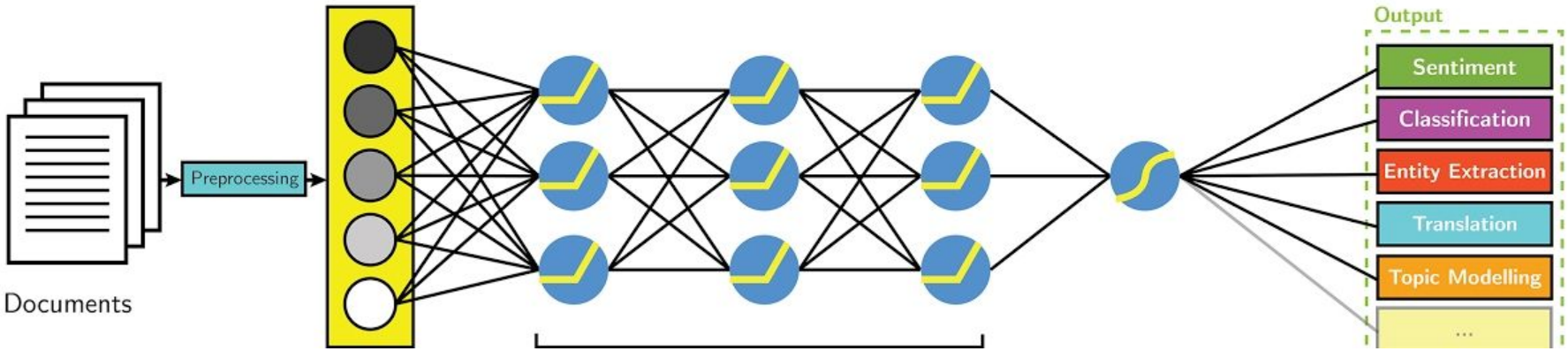
Anos 2010 – Deep Neural network

- Uso de placas gráficas implementando redes neurais com várias camadas ocultas e com capacidade de extrair diversas *features*



LINGUÍSTICA COMPUTACIONAL





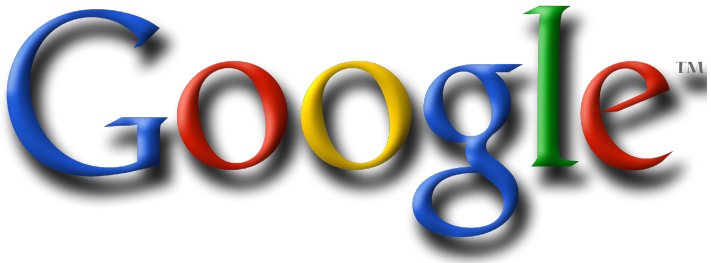
LINGUÍSTICA COMPUTACIONAL

Por que pesquisar em LC?

LINGUÍSTICA COMPUTACIONAL

Por que pesquisar em LC?

- *Porque existem grandes oportunidades para pesquisa e desenvolvimento de aplicações úteis.*
- *Gigantes de TI pesquisam e contratam nessa área.*



LINGUÍSTICA COMPUTACIONAL

Por que pesquisar em LC?

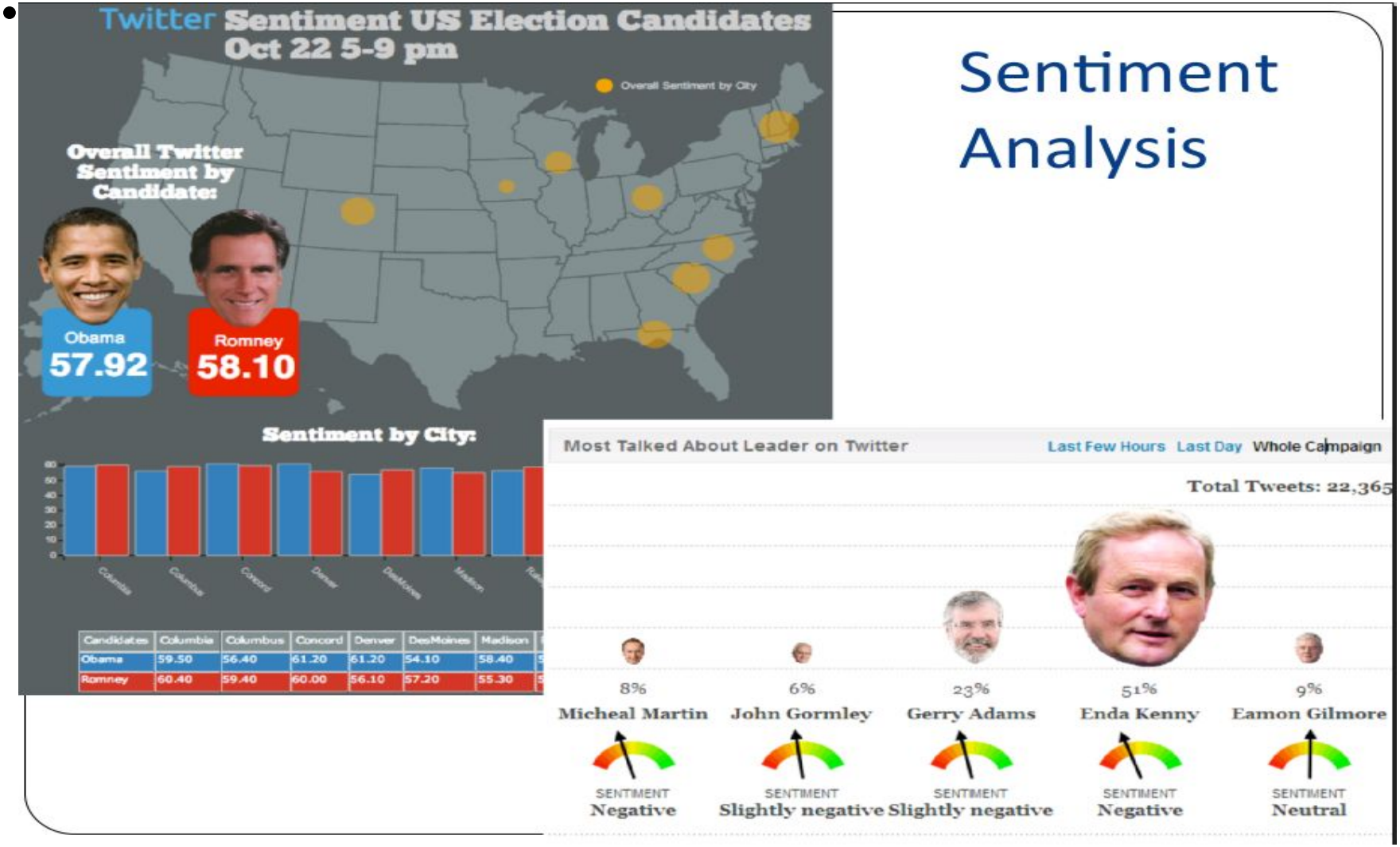
- *Desenvolvimento de sistemas de perguntas e respostas.*
-

WILLIAM WILKINSON'S "AN ACCOUNT OF THE PRINCIPALITIES OF WALLACHIA AND MOLDOVIA" INSPIRED THIS AUTHOR'S MOST FAMOUS NOVEL

Bram Stoker

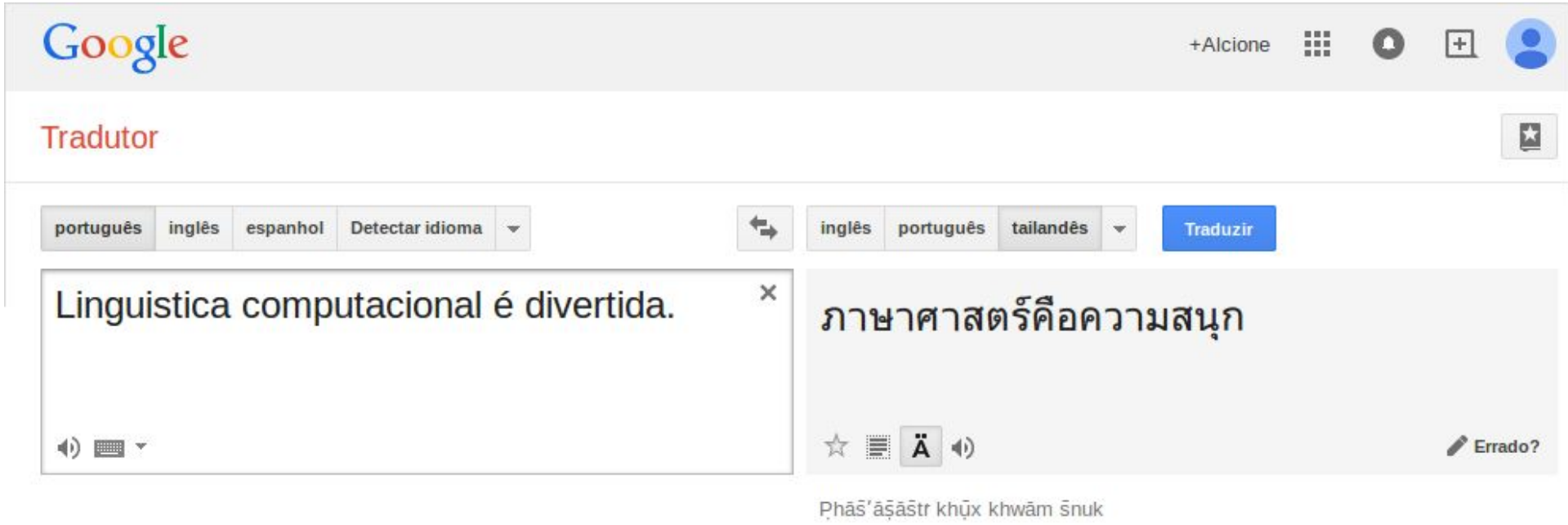
LINGUÍSTICA COMPUTACIONAL

Por que pesquisar em LC? *Análise de sentimento*

- 

LINGUÍSTICA COMPUTACIONAL

Por que pesquisar em LC? *Tradução automática*

- 

LINGUÍSTICA COMPUTACIONAL

Por que pesquisar em LC?

Extração de informação, Query/answer, Classificação de texto, atribuição de autoria, conversação, sumarização, e muito mais ...

-

LINGUÍSTICA COMPUTACIONAL

Áreas de pesquisa

Resolvida em parte

Detecção de spam

Let's go to Agra!

Buy V1AGRA …

Anotação Part-of-speech (POS)

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in NY

Progredindo

Análise de sentimento

Melhor pizza de Juiz de Fora!

O garçon nos ignorou por 20 min.

Resolução de coreferência

Carter told Mubarak he shouldn't run again.

Word sense disambiguation

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

Ainda difícil

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Sumarização

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

LINGUÍSTICA COMPUTACIONAL

Base teórica

LINGUÍSTICA COMPUTACIONAL

Conhecimentos úteis:

- Probabilidade básica
- Estatística básica
- Álgebra linear básica
- Cálculo
- Aprendizado de máquina
- IA
- Algoritmos
- Linguística

LINGUÍSTICA COMPUTACIONAL

Modelos:

- Autômatos finitos
- Modelos de Markov
- Modelos de alinhamento
- Modelos de Espaço Vetorial (IR)
- Modelos de Rede

LINGUÍSTICA COMPUTACIONAL

Programação dinâmica:

- Mínima distância de edição
- Algoritmo de Viterbi
- Baum-Welch/Forward-Backward

LINGUÍSTICA COMPUTACIONAL

Classificadores baseados em aprendizado de máquina

Classificadores Simples:

- Naïve Bayes
- Logistic Regression (MaxEnt)
- Decision Trees
- Perceptron
- MLP

Modelos para sequências:

- Hidden Markov Models
- Maximum Entropy Markov Models
- Conditional Random Fields
- RNN

Abordagem Tradicional

PROCESSAMENTO DA LINGUAGEM NATURAL

Etapas do Processo de compreensão da LN

- Análise Morfológica
- Análise Sintática
- Análise Semântica
- Integração de Discurso
- Análise Pragmática

PROCESSAMENTO DA LINGUAGEM NATURAL

Terminologia

Sintagma – grupo de elementos linguísticos classificados de acordo com a categoria sintática de seu elemento núcleo:

Sintagma nominal (SN) – núcleo substantivo.

Sintagma verbal (SV) – núcleo verbo.

Sintagma preposicional (SP) – núcleo preposição.

Sintagma adverbial (SAdv) – núcleo advérbio o locução adverbial.

PROCESSAMENTO DA LINGUAGEM NATURAL

Terminologia

Os sintagmas são unidades linguísticas de nível intermediário, sendo constituintes de uma unidade de nível superior, a frase.

SN – João; o menino; a maçã verde; o gato de rabo longo.
SV – chove; chegou cedo; tem estado doente; falaram de Maria Maria a Pedro.
SP – para você; de Maria a Pedro.
Sadv – Cedo; muito rapidamente.

PROCESSAMENTO DA LINGUAGEM NATURAL

Análise Morfológica

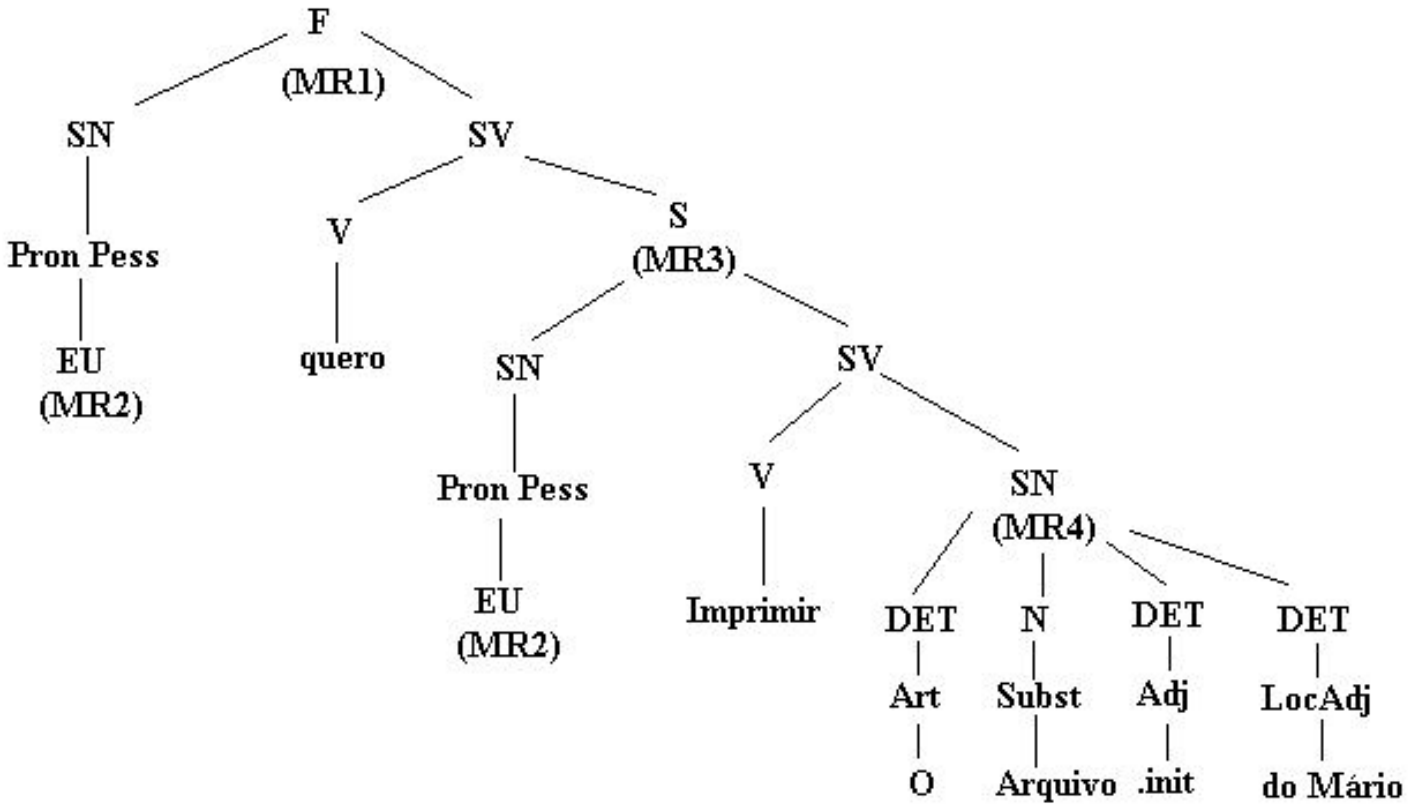
É a análise da forma e inflexão das palavras.
exemplo: *Eu quero imprimir o arquivo .init do Mário.*

Eu - pronome
Quero - Verbo querer, presente, 1a. pessoa do singular.
imprimir - Verbo imprimir, infinitivo
o - artigo
arquivo - substantivo
.init - adjetivo
do - contração da preposição de e do artigo o
Mário - nome prório

PROCESSAMENTO DA LINGUAGEM NATURAL

Análise Sintática

Identifica a estrutura da frase por meio das regras de sintaxe da língua. Como resultado retorna uma árvore que reflete a estrutura.



Análise da sentença: *Eu quero imprimir o arquivo .init do Mário.*

PROCESSAMENTO DA LINGUAGEM NATURAL

Análise Semântica

Os elementos da estrutura sintática são mapeados para uma estrutura de representação de conhecimento.

MR1		(toda frase)
Instância:	Desejo	
Agente:	MR2	(eu)
Objeto:	MR3	(evento de impressão)
MR2		(eu)
MR3		
Instância:	Impressão	
Agente:	MR2	
Objeto:	MR4	(arquivo .init do Mario
MR4		
Instância:	Estrut-arq	
extensão:	.init	
dono:	MR5	(Mário)
MR5		
Instância:	Pessoa	(mário)
nome:	Mário	

PROCESSAMENTO DA LINGUAGEM NATURAL

Integração de Discurso

O significado capturado da frase é integrado no contexto. Neste momento referências por meio de pronomes são resolvidas.

No caso do exemplo é preciso resolver a quem se refere o pronome "eu" e quem é "Mario".

Análise Pragmática

Interpreta o que realmente se quis dizer: "!Você sabe que horas são?"

No caso do exemplo o significado e um comando de impressão do arquivo do tipo lpr /home/mario/.init

Deep Learning para PLN

Word Embedding

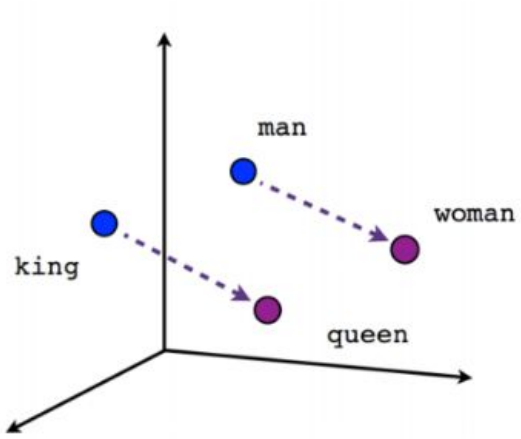
Introdução

- A Word Embedding é o nome de um conjunto de técnicas de modelagem de linguagem e de aprendizado de recursos no processamento de linguagem natural (NLP), em que palavras ou frases do vocabulário são mapeadas para vetores de números reais.

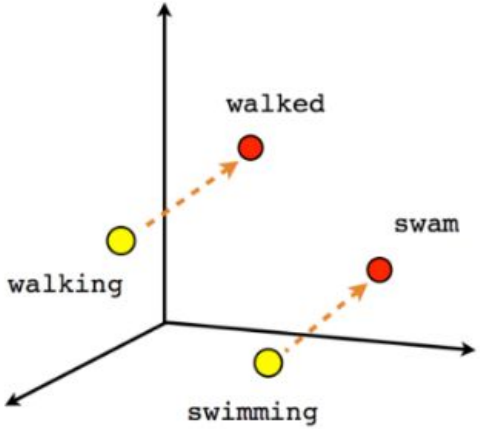
- Conceitualmente, envolve uma incorporação matemática de um espaço com uma dimensão por palavra para um espaço vetorial contínuo com uma dimensão muito maior.

Word Embedding

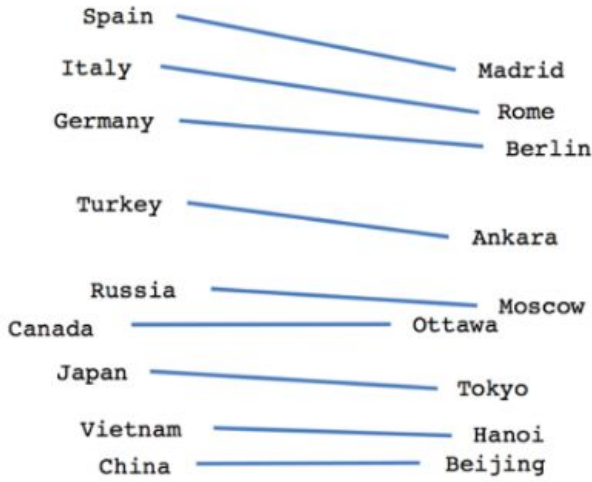
Introdução



Male-Female



Verb tense



Country-Capital

vetor[Queen] = vetor[king] - vetor[Man] + vetor[Woman]

Word Embedding

Introdução

É possível obter bastante informação representando o uma palavra em termos de seus "vizinhos".

"You shall know a word by the company it keeps"

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

Word Embedding

Introdução

O significado das palavras é definido em termo de vetores.

É construído um vetor denso para cada palavra, feito de tal forma a ser útil em prever as outras palavras que aparecem em seu contexto.

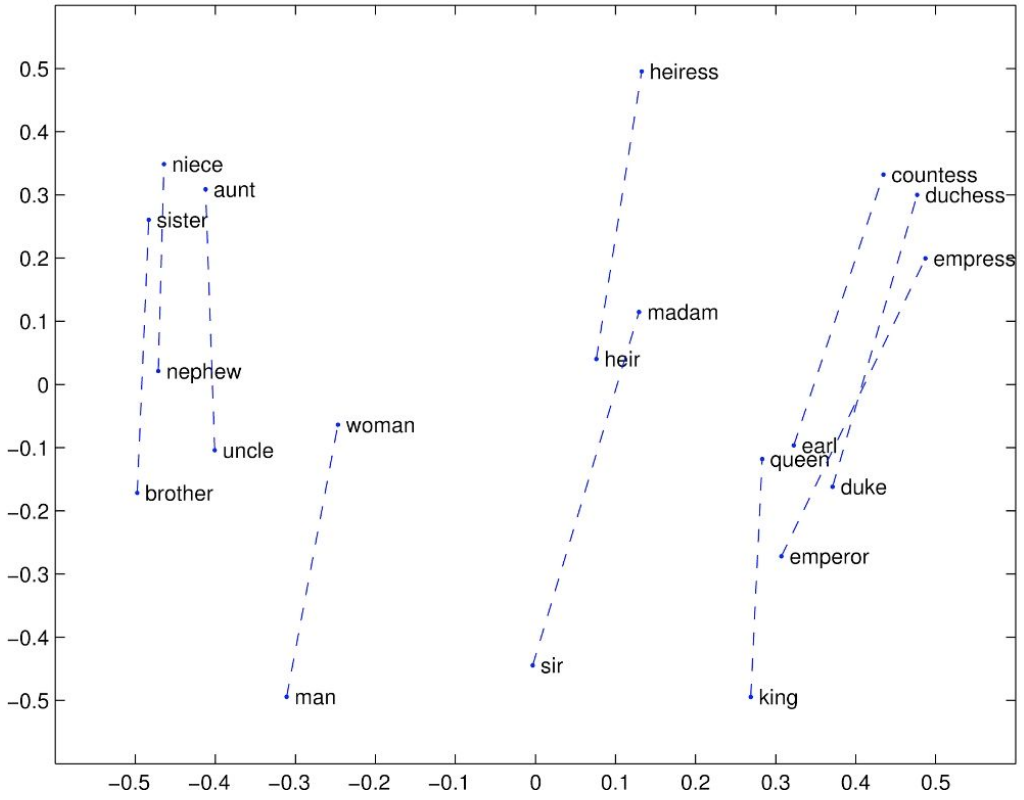
$$\textit{linguistics} = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Word Embedding

Glove

Resultados do Glove

- Visualizações



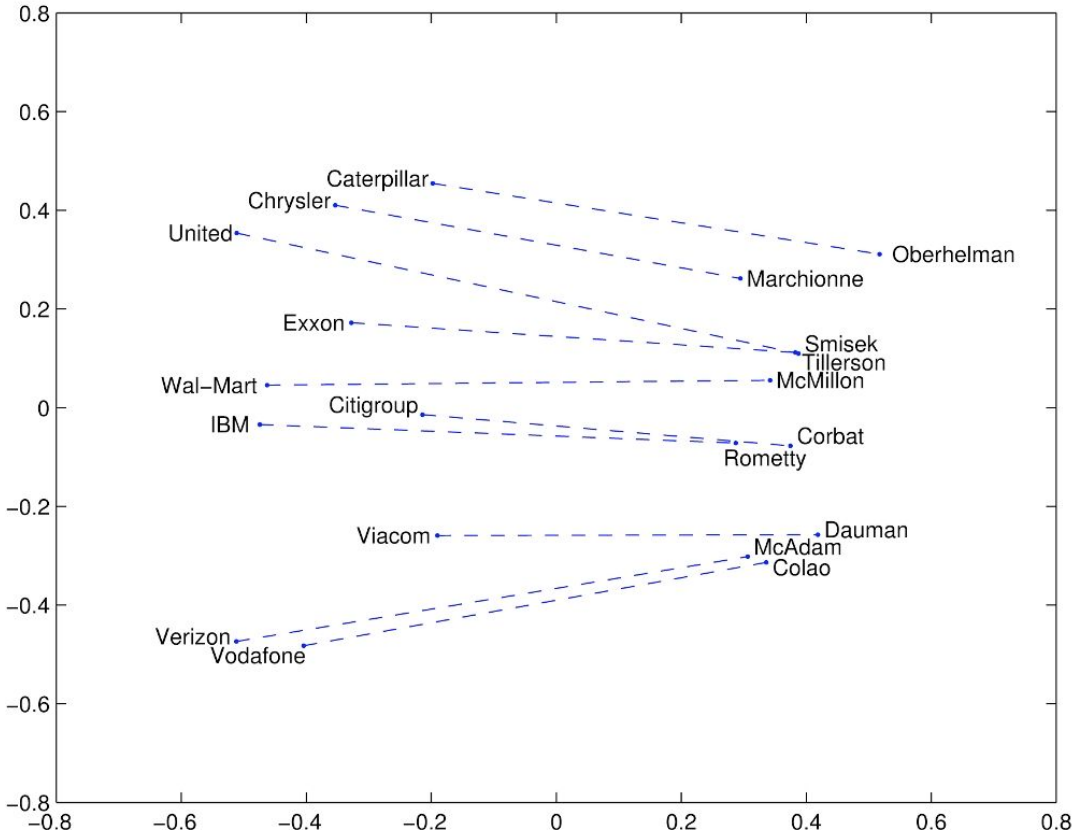
Word Embedding

Glove

Resultados do Glove

- Visualizações

Presidentes de empresas

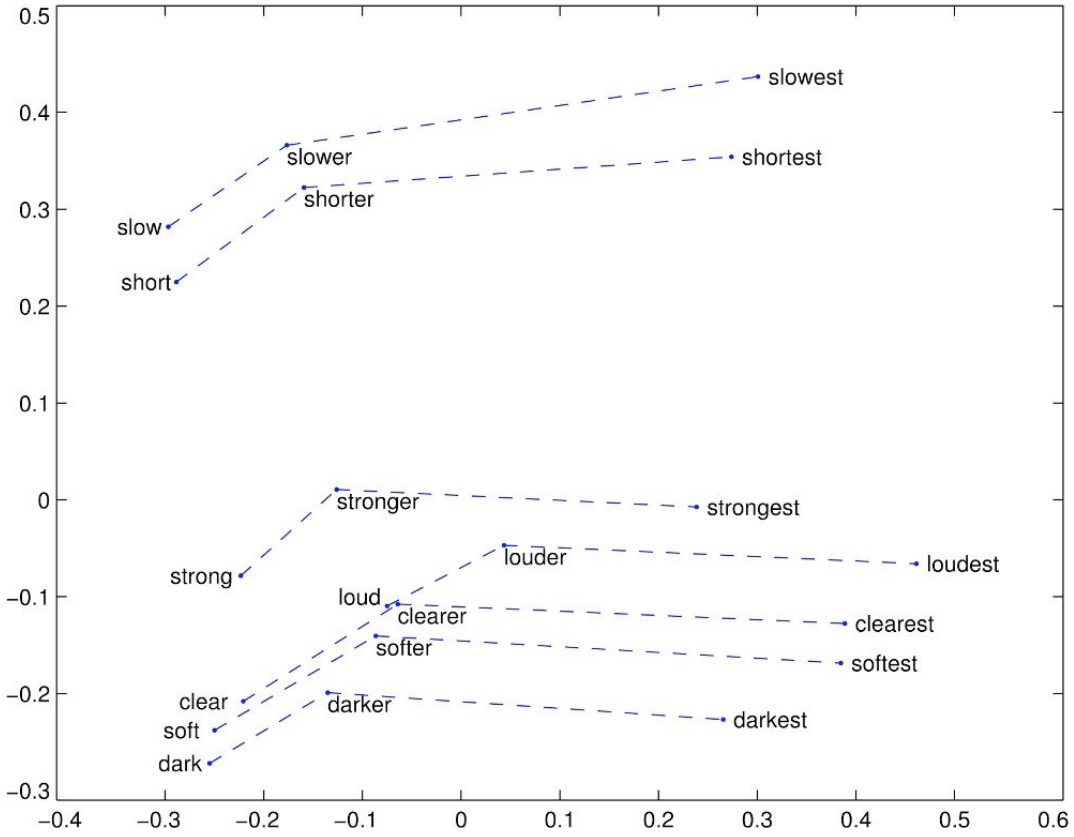


Word Embedding

Glove

Resultados do Glove

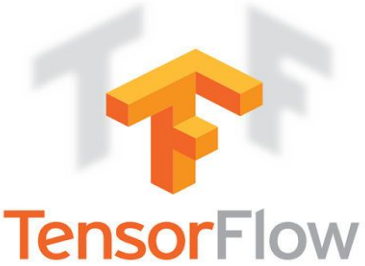
- Visualizações Superlativos



LINGUÍSTICA COMPUTACIONAL

Ferramentas

LINGUÍSTICA COMPUTACIONAL

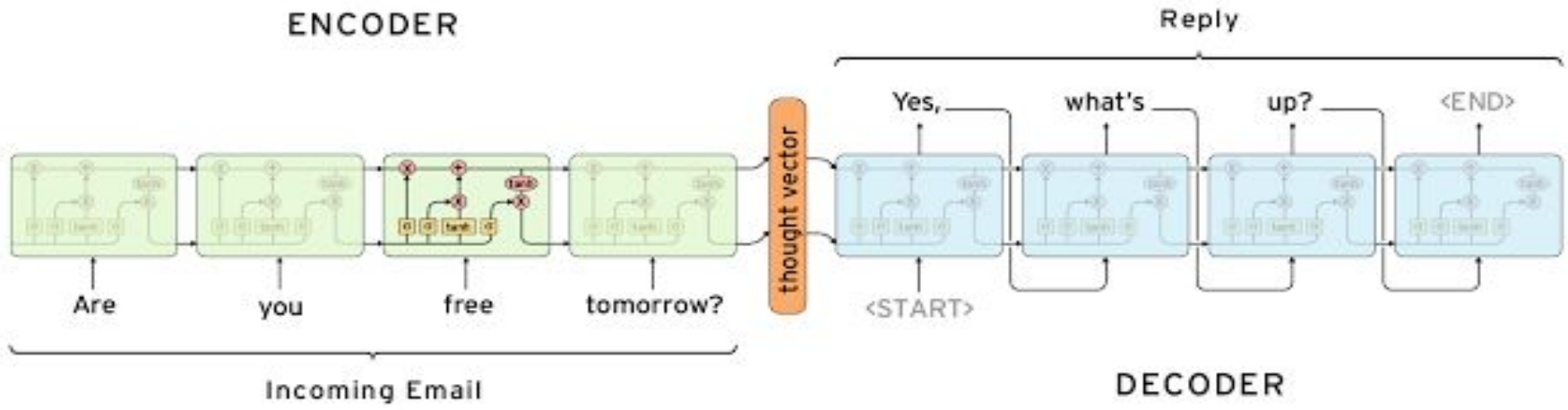


TensorFlow ™ é uma biblioteca de software de código aberto para computação numérica usando gráficos de fluxo de dados. Os nós no gráfico representam operações matemáticas, enquanto os vértices do gráfico representam as matrizes de dados multidimensionais (tensores) que são comunicados entre eles.
A arquitetura flexível permite computação para uma ou mais CPUs ou GPUs em um desktop, servidor ou dispositivo móvel com uma única API.

LINGUÍSTICA COMPUTACIONAL

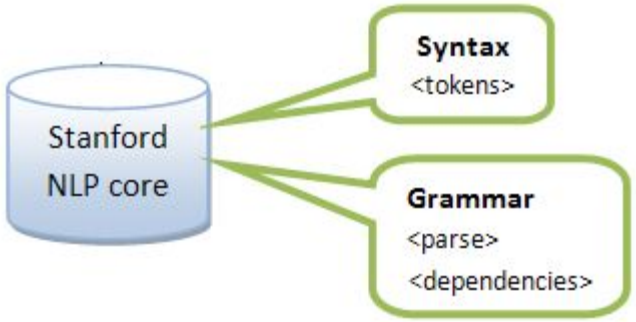
TensorFlow

Arquiteturas de Aprendizagem Profunda como sequência para sequência são especialmente adequadas para a geração de texto. No entanto, ainda estamos nos estágios iniciais de construção de modelos generativos que funcionam razoavelmente bem.



LINGUÍSTICA COMPUTACIONAL

Ferramentas



Natural Language Analyses with NLTK

spaCy

openNLP™

LINGUÍSTICA COMPUTACIONAL

FIM