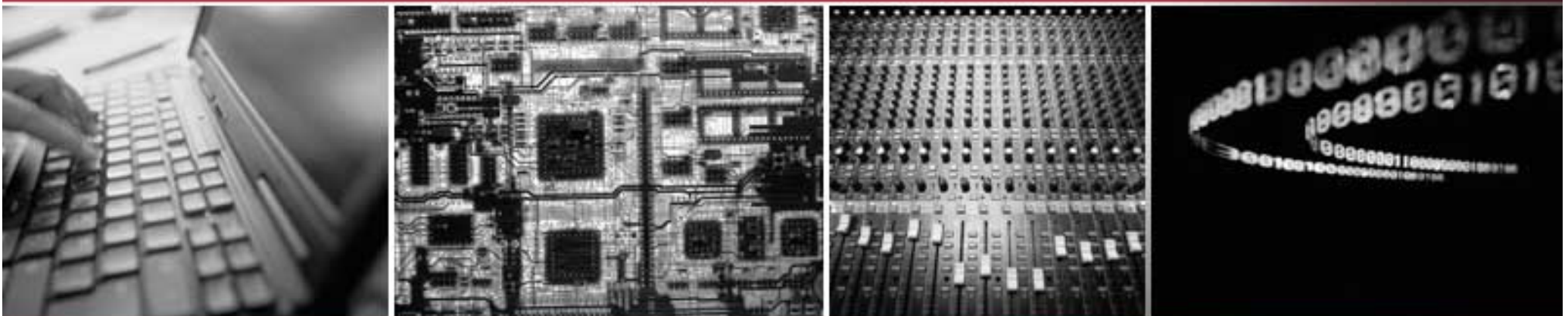


Fundamentos de Processamento de Linguagem Natural (PLN)

Prof. Dr. Emerson Cabrera Paraiso



Pontifícia Universidade Católica do Paraná (PUCPR)
Programa de Pós-Graduação em Informática (PPGIa)

Mestrado e Doutorado
www.ppgia.pucpr.br

Apresentação do Professor

- Prof. Dr. Emerson Cabrera Paraiso
 - Graduação
 - Curso: Engenharia de Computação
 - Instituição: PUCPR
 - Mestrado
 - Título: Mestrado em Engenharia Elétrica e Informática Industrial
 - Instituição: UTFPR (antigo CEFET-PR)
 - Dissertação: Concepção e Implementação de um Sistema Multi-agentes para Monitoração de Processos Industriais
 - Doutorado
 - Título: Doutorado em Sistemas de Informação
 - Instituição: Université de Technologie de Compiègne – France
 - Tese: Une Interface Conversationnelle pour une Aide Intelligente

#2



Apresentação do Professor (cont.)

- Atividades na PUCPR
 - Graduação
 - Raciocínio Algorítmico – BSI
 - Interação Humano-Computador - BES
 - Pós-Graduação
 - Coordenador do Programa de Pós-Graduação em Informática – PPGIa
 - www.ppgia.pucpr.br
 - Grupo de Pesquisa Descoberta de Conhecimento e Aprendizagem de Máquina.



Detalhes do Curso

- **Identificação**
 - Fundamentos de Processamento de Linguagem Natural (PLN)
- **Objetivos**
 - Apresentar os fundamentos do Processamento de Linguagem Natural, da Recuperação da Informação e Mineração de Textos.
- **Módulo de Sequência**
 - Aplicações de Processamento de Linguagem Natural (1º sem 2021)

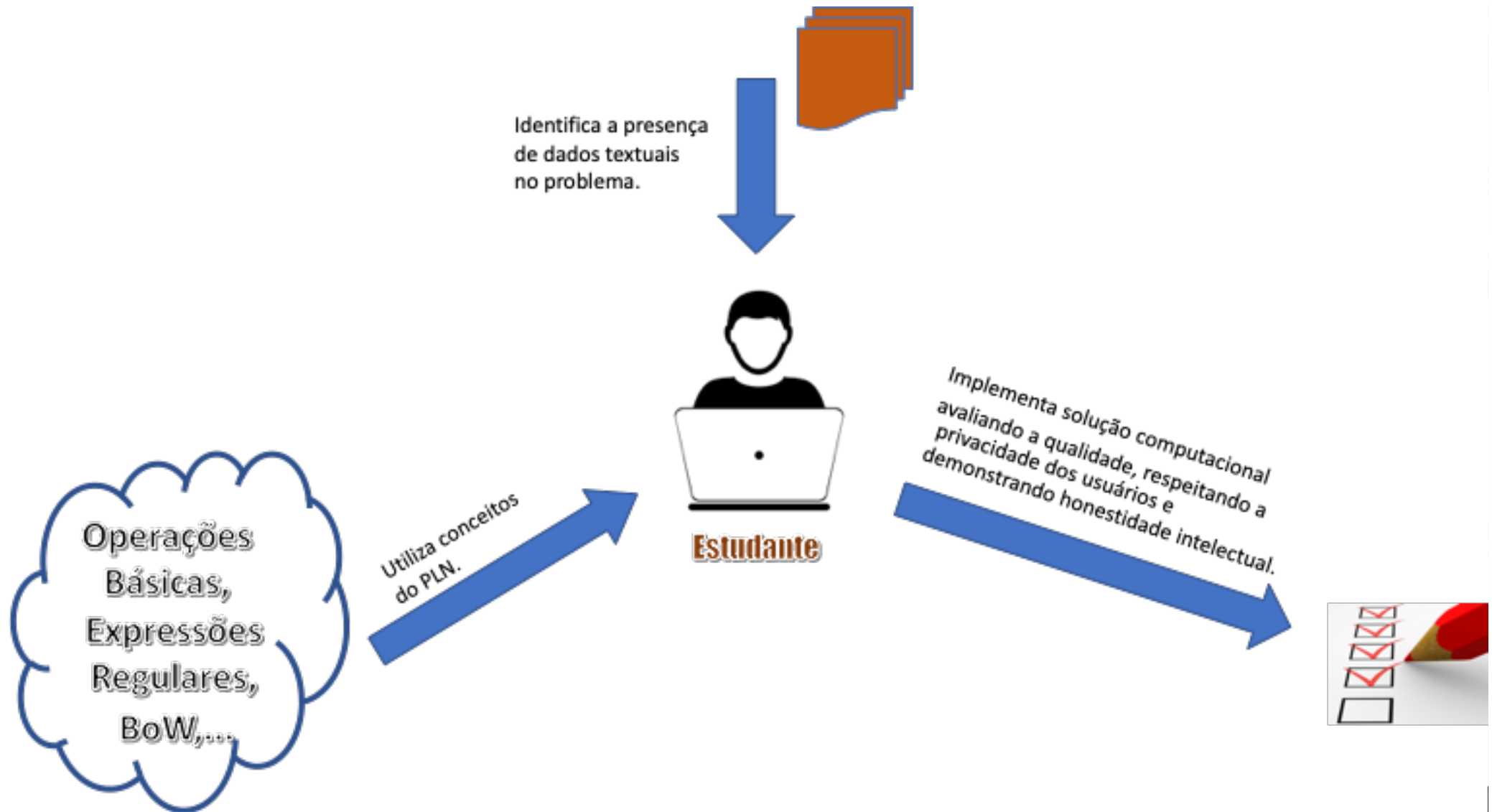


Temas de Estudo

- Conceitos básicos:
 - Processamento de Linguagem Natural, Recuperação da Informação, Linguística Computacional
- Processamento básico de texto:
 - Expressões Regulares, Similaridade entre palavras e textos, etc.
- Recursos Léxicos:
 - Ontologia Léxica, Word Embeddings
- Extração e Recuperação da Informação
- Mineração de Textos:
 - Classificação
 - Análise de Sentimentos



Mapa Mental



Calendário de Aulas

- Calendário (pequenas adaptações podem ocorrer):
 - 09/05 - Apresentação/Introdução à disciplina
 - 23/05 - Aula
 - 06/06 – Aula
 - 20/06 – Aula
 - 04/07 – Aula
 - 18/07 – Apresentação de trabalhos



Metodologia de Trabalho

- Aulas práticas para a realização de exercícios.
- Material didático de apoio sob a forma de PDFs.
- Desenvolvimento de trabalhos a serem especificados para avaliação do rendimento do estudante.



Algumas Referências

- Notas de aula.
- Natural Language Processing with Python– Analyzing Text with the Natural Language Toolkit. Steven Bird, Ewan Klein, and Edward Loper. (disponível em: <http://www.nltk.org/book/>)
- Text Mining: Predictive methods for Analyzing Unstructured Information. Sholon Weiss, Nitin Induskhya, Tong Zhang, and Fred J. Damerau.
- M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130–137.
- Notas de aula do prof. Dan Jurafski (Stanford)
- Site de recursos linguísticos para o Português: www.linguateca.pt
- ACL Anthology: <http://aclweb.org/anthology/>

- Todo contato deve ser feito preferencialmente via mensagens do Blackboard.

Política de Direitos Autorais

- Todo e qualquer artefato produzido pelos alunos poderá ser disponibilizado para acesso aberto.
- A produção de cada estudante será corrigida e, na indicação de cópia de material de terceiros, sem a devida referência de autoria, levará a atribuição da nota zero.
- Atenção: cópia é crime e não será tolerada nesta disciplina.



Avaliações

| Atividade | Datas | Peso |
|------------------------------|------------------|------|
| Trabalhos práticos pontuais | Durante as aulas | 60% |
| Trabalho final da disciplina | 18/07/2020 | 40% |



Sumário – Primeira Aula

- Contextualização
- Primeiras Definições
- Exercício Inicial

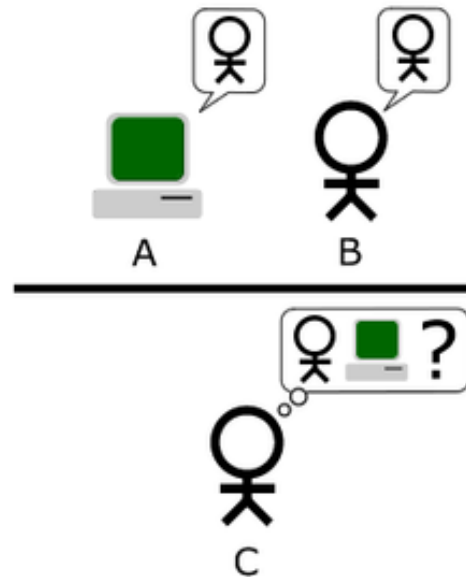


Reflexão: o que é uma máquina inteligente para você?

A computer would deserve to be called intelligent if it could deceive a human into believing that it was human.

Alan Turing

Teste de Turing



Extraído de: https://pt.wikipedia.org/wiki/Teste_de_Turing#CITEREFSaygin2000

**I'M NOT SCARED OF A COMPUTER
PASSING THE TURING TEST...**

**I'M TERRIFIED OF ONE THAT
INTENTIONALLY FAILS IT**

memegenerator.net


#16



Contextualização

- Permitir que uma máquina interprete um texto em linguagem natural é sem dúvida um dos maiores desafios da computação:
 - Textos em linguagem natural podem ser ambíguos, subjetivos, conter erros.
 - “A menina disse à colega que sua mãe havia chegado.”
 - “A vaca se diverte com a pata na lama.”
 - “Pode deixar, darei um geito.”
 - Neologismos:
 - “Retweet”
- Trata-se de uma área de pesquisa interdisciplinar.
- Uma das áreas de maior atenção “comercial” dos últimos anos.

Um Pouco de História

- O PLN começou a se desenvolver no início dos anos 1950.
- A primeira tarefa que chamou atenção foi a tradução automática:
 - Russo  Inglês
- Na década de 1960, Joseph Weizenbaum desenvolveu o ELIZA. ELIZA simula a conversação entre um humano e um computador, tentando “dar a impressão” ao humano de que entende o que este fala (no caso escreve).
- A partir dos anos 1980, sistemas baseados em regras começaram a proliferar.
- Surgem os parsers e as ontologias.
- Um grande passo para a evolução da área é dado com o desenvolvimento do Aprendizado de Máquina (Machine Learning).

Mais detalhes em: https://en.wikipedia.org/wiki/History_of_natural_language_processing

Por que o interesse recente?

- Há muita informação textual (dado não estruturado) acumulada na Web, nas empresas, nos computadores das pessoas.

The Amazon logo, featuring the word "amazon" in a black, lowercase, sans-serif font, with a curved orange arrow underneath it pointing from the 'a' to the 'z'.The Google logo, consisting of the word "Google" in its multi-colored, rounded, sans-serif font.The Facebook logo, featuring the word "facebook" in white, lowercase, sans-serif font, centered within a blue rectangular background.The WhatsApp logo, featuring a green speech bubble with a white telephone handset icon inside, followed by the word "WhatsApp" in a bold, dark green, sans-serif font.The Twitter logo, featuring the word "twitter" in white, lowercase, sans-serif font, followed by a white bird icon, all on a blue rectangular background.

Onde pode ter PLN aqui?



#19



Problemas Ocorrem!



- Das mais simples:
 - Busca por palavra-chave
 - Identificação de sinônimos
 - Verificação da escrita (ortografia)
 - Extração da informação
- As mais sofisticadas:
 - Tradução automática
 - Reconhecimento e geração da fala
 - Sistemas de diálogo e Chatbots

Algumas Aplicações

- Recuperação de informação textual:
 - 6.586.013.574 buscas na web todo dia (estimativa de 2017)



Algumas Aplicações (cont.)

- Extração da informação a partir de dados textuais:

I_1 - Bom dia.

I_2 - Bom dia.

I_1 - Gostaria de uma informação.

I_2 - Pois não, pode perguntar.

I_1 - De quanto tempo é o estágio probatório?

I_2 - O estágio probatório é de 3 anos contados a partir da data de posse.

I_1 - Obrigado

I_2 - Sem problemas.

O diálogo tem um domínio específico!

Algumas Aplicações (cont.)

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg

Date: Jan-16-2012

Start: 10:00am

End: 11:30am

Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Create new Calendar entry

Notas de aula: Dan Jurafski

#24



Mineração de Opiniões e Análise de Sentimentos

●●●●● Avaliou em 2 semanas atrás

Conforto e localização

O hotel esta localizado no coração de Blumenau, facil acesso a pé para diversos pontos. O quarta é bem confortavel, limpo e aconchegante. Cofre e secador disponivel no quarto. Mas o destaque fica para o café da manha que é um show a parte. Além... [Mais](#)

Tyler Adams @TheSlackerMcFly

Replying to @EddieTrunk

LOVED #TrunkFest on @AXSTV @AXSTVConcerts! Welcome back to tv with your own show Eddie! Sure hope it got huge ratings so that @mcuban might wanna bring @ThatMetalShow back 😊! Can't wait for next weeks show! Have a good vacation this week Ed!

Linguagem “natural”



#25



Sistemas de Recomendação

Customers Who Bought This Item Also Bought



A Curious History of Food and Drink
› Ian Crofton
★★★★☆ 11
Hardcover
\$15.06 ✓Prime



Consider the Fork: A History of How We Cook and Eat
› Bee Wilson
★★★★☆ 217
Paperback
\$11.28 ✓Prime



Fifty Foods That Changed the Course of History (Fifty Things That Changed the...)
Bill Price
★★★★☆ 2
Hardcover
\$23.10 ✓Prime

Tradução

Google

Tradutor Desativar tradução instantânea

Francês Português Inglês Detectar idioma

Português Francês Japonês Traduzir

Processamento de linguagem natural

自然言語処理

34/5000

Shizen gengo shori

Sugerir uma edição

#27



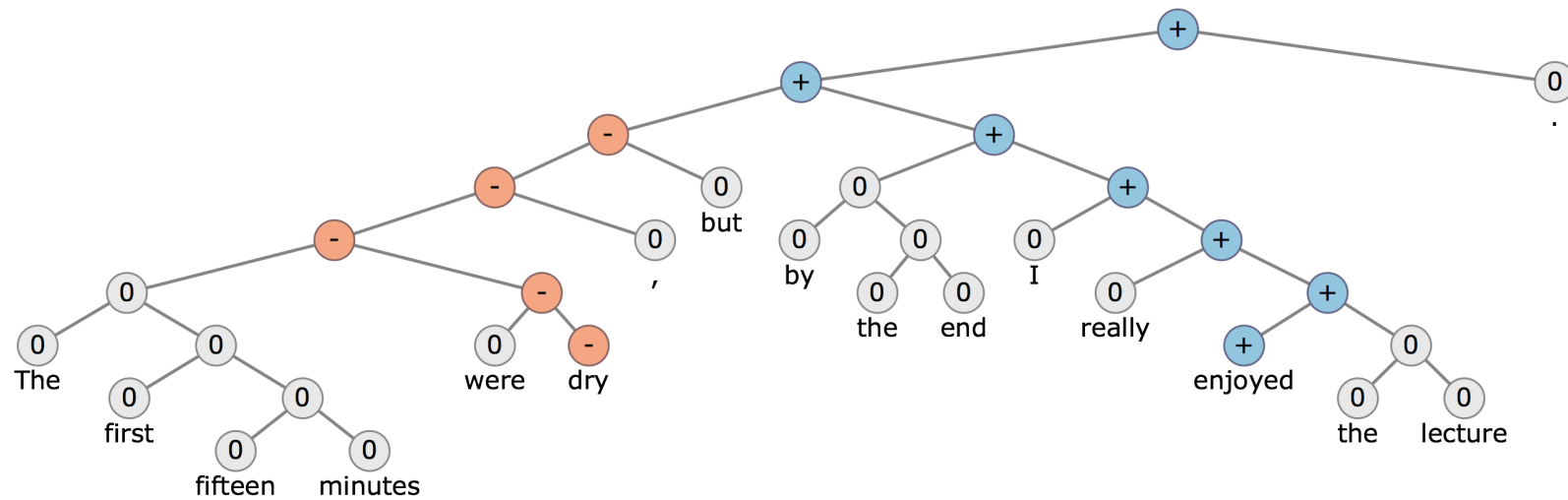
Programa de Pós-Graduação em Informática (PPGIA)

www.ppgia.pucpr.br

Copyright©2020 – Prof. Dr. Emerson Cabrera Paraiso. Todos os direitos reservados.

Mestrado e Doutorado

Parsing



Extraído de: <http://web.stanford.edu/class/cs224n/>

mostly solved

Spam detection

Let's go to Agra!



Buy VIAGRA ...



Part-of-speech (POS) tagging

ADJ ADJ NOUN VERB ADV

Colorless green ideas sleep furiously.

Named entity recognition (NER)

PERSON ORG LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!



The waiter ignored us for 20 minutes.



Coreference resolution

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

still really hard

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

Summarization

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?



Notas de aula: Dan Jurafski

Desafios para o Processamento do Português

- Recursos mais limitados
 - Parser, part-of-speech, ...
 - Ontologias, dicionários
 - Brasileiro (PT-BR)
 - Europeu (PT-EU)
 - Reconhecimento da Fala
 - Corpora

Identificando o Dado Textual

- Dado textual = não estruturado
- Características do dado textual:
 - Não tem tipo (como no dado estruturado – tabela em um BD);
 - Disponível em texto-puro (ASCII ou UNICODE).

Extração do Dado Textual

- Origem distintas:
 - Web (html)
 - Redes Sociais: posts
 - Sistemas de Informação: nome de pessoas, endereço eletrônico, ...
- Exemplo para arquivo .html:
 - texto “espalhado” ao longo do arquivo .html e suas tags.
 - conjunto de funções para extração (parser).
 - Veja isto:
<https://pythonhelp.wordpress.com/2013/03/18/webscraping-em-python/>

Quiz

- Vamos responder o Quiz sobre dado textual disponível no BB.



Respostas

P1) O dado textual também é conhecido como:

Dado não estruturado

P2) O código-fonte escrito em Python é um dado textual?

Verdadeiro

P3) O nome completo de uma pessoa, gravado em um banco de dados relacional, não é um dado textual?

Falso

P4) O campo “endereço” de um formulário de cadastro de um candidato ao vestibular, é um campo textual?

Verdadeiro

Respostas (cont.)

P5) A foto a seguir, pode ser considerada um dado textual?



Falso

Compromisso Ético

- Somente dados públicos e disponibilizados com a autorização de seus proprietários podem ser utilizados.
- Compromisso com o respeito à Lei Geral de Proteção de Dados Pessoais (LGPD).
- Dê uma olhada em: <https://www.serpro.gov.br/lgpd>.



Conceitos Básicos

- Linguagem natural: linguagens que são utilizadas para comunicação do dia a dia por humanos (português brasileiro, português europeu, inglês, ...).
- Processamento de Linguagem Natural (PLN): qualquer manipulação computacional de linguagens naturais. De contagem de palavras à compreensão semântica.
- Linguística Computacional: associada à PLN, estuda os fenômenos linguísticos para apoiar o computador na interpretação e geração da linguagem natural.

Conceitos Básicos (cont.)

- Corpus: conjunto de textos, normalmente normalizados e rotulados.
- Corpora: conjunto de Corpus.
- Entidade Nomeada: são expressões que nomeiam pessoas, organizações, locais, tempos e quantidades.
 - Exemplo: “São Paulo”, “Brasil”, “Pedro Alvares Cabral”, “ONU”, etc.
 - Dificuldades: “SP”, “S.P.”, “S. Paulo”, “São Paulo”, ...

Conceitos Básicos (cont.)

- Léxico: conjunto de palavras de um dado idioma.
 - O léxico de uma língua não é “fechado” ou fixo.
 - Podem influenciar no léxico:
 - Nomes próprios;
 - Abreviações e siglas;
 - Gírias, etc.



Exercício Inicial

- Implemente um algoritmo em Python para resolver o seguinte problema:
- Dado o seguinte léxico:
[abacate, abacaxi, abobora, abobrinha, ananás, maçã, mamão, manga, melancia, melão, mexerica, morango]
- Indicar a palavra mais “próxima”:

abacati

abacate
abacaxi
abobora
abobrinha

Desafios: o que é “similar”
neste contexto?
Como medir o grau de
“similaridade” entre palavras?

Similaridade Sintática

- A similaridade sintática entre strings pode ser medida por uma função de distância.
- São muito utilizadas as distâncias de Hamming e a de Levenshtein (Edit Distance).
- A distância de edição (Edit Distance) é definida pelo número de inserções, exclusões e substituições realizadas na comparação entre as strings envolvidas.
- Exemplos:
 - “color” -> “colour”: ED = 1
 - “survey” -> “surgery”: ED = 2

Cálculo do N-Gram

- Um N-gram pode ser entendido como um conjunto de “gramas” consecutivos, onde um “grama” pode ser uma letra ou palavra.
- O n-gram é muito útil em diversas tarefas do PLN.
- Exemplo:
 - calcular o grau de similaridade sintática entre as seguintes palavras: “parar” e “parado”
 - inicialmente devemos definir o valor de N: $N = 2$ (digrama)
 - “parar” = {pa, ar, ra, ar} (4 digramas e 2 únicos: pa, ra)
 - “parado” = {pa, ar, ra, ad, do} (5 digramas e 5 únicos: pa, ar, ra, ad, do)
 - Para o cálculo da similaridade, usar a fórmula:
 - $S = 2C / A + B$
 - Onde:
 - » A é o número de n-gramas únicos na primeira palavra
 - » B é o número de n-gramas únicos na segunda palavra
 - » C é o número de digramas únicos compartilhados
 - $S = 2 * 2 / 2 + 5 = 0.58$

Outros Exemplos

- P1 = “parana” e P2 = “paranaense”
 - {pa, ar, ra, an, na}: únicos = {pa, ar, ra, an, na}
 - {pa, ar, ra, an, na, ae, en, ns, se} : únicos = {pa, ar, ra, an, na, ae, en, ns, se}
 - Compartilhados = {pa, ar, ra, an, na}
 - $S = 2 * 5 / 5 + 9 = 0,71$
- P1 = “carro” e P2 = “aviao”
 - {ca, ar, rr, ro}: únicos = {ca, ar, rr, ro}
 - {av, vi, ia, ao} : únicos = {av, vi, ia, ao}
 - Compartilhados = {0}
 - $S = 2 * 0 / 4 + 4 = 0$

Código em Python

- Ao final entregar o código implementado em atividade criada no BB: Aula 1 - Similaridade Sintática.
- Avaliar diferentes *thresholds* de distância.
- O exercício pode ser feito em dupla.

