



PROCESSAMENTO DE LINGUAGEM NATURAL

Prof. Thiago A. S. Pardo



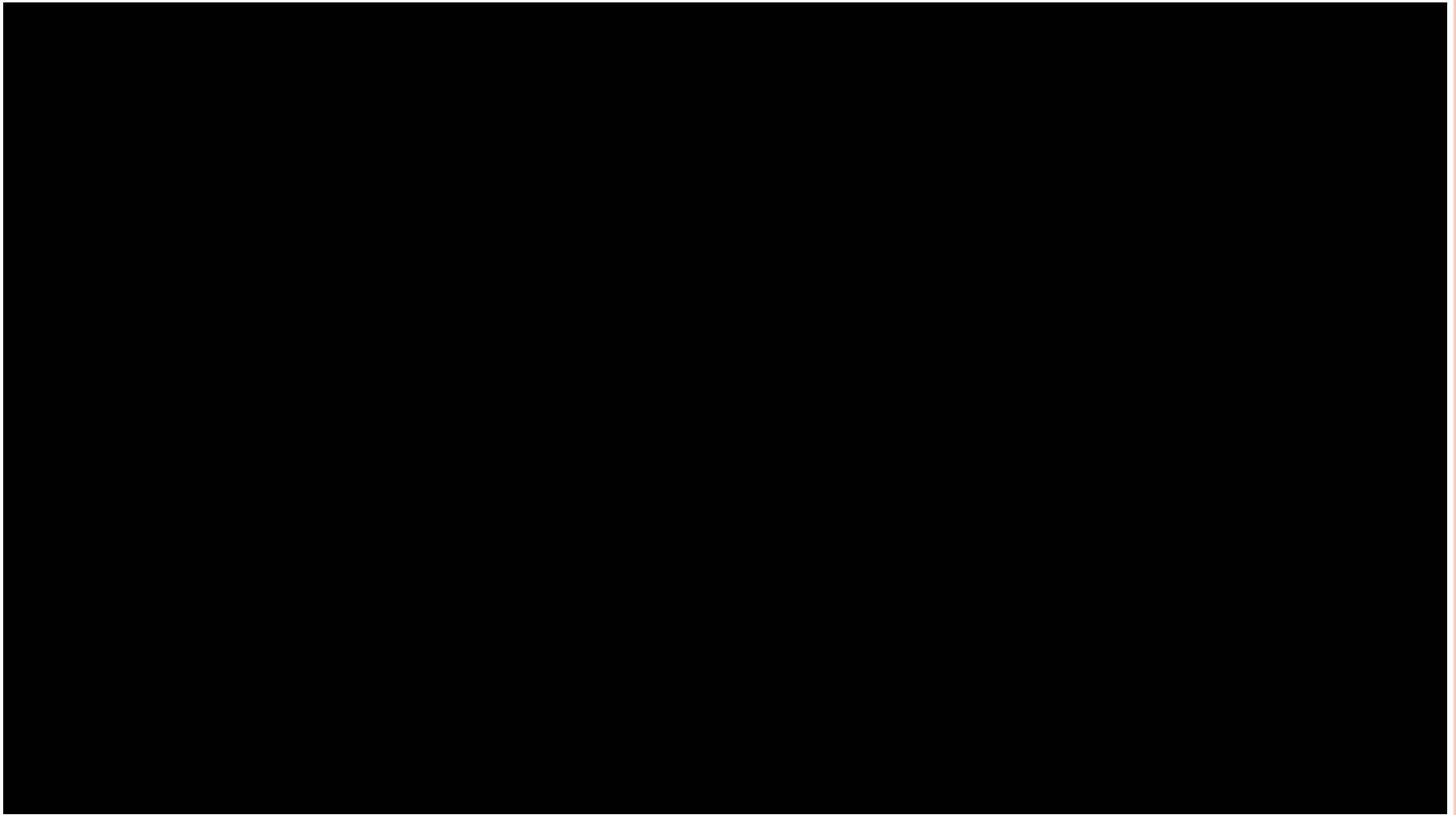


Pictogramas: linguagem!

RUMO ÀS ESTRELAS



DE PICTOGRAMAS À LÍNGUA ALIENÍGENA!



O CINEMA COMO INSPIRAÇÃO

- Prometheus (2012), de Ridley Scott



DIVERSAS REFERÊNCIAS

- Jornada nas Estrelas
- Guerra nas Estrelas
- IA
- Matrix
- Eu, robô
- O homem bicentenário
- Wall-E
- Ela
- Ex-Machina: Instinto Artificial
- Homem de Ferro
- Etc.



LÍNGUA NATURAL

- Língua humana



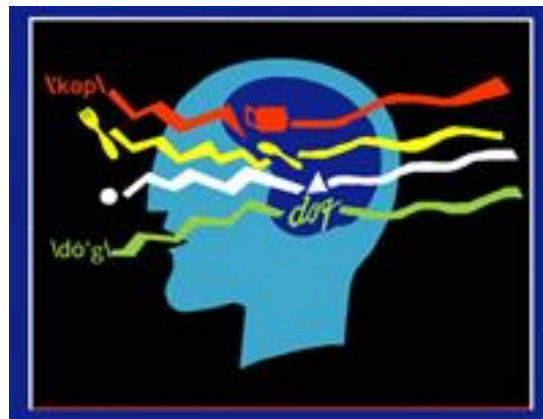
- Em oposição às linguagens artificiais

- Matemática, lógica, linguagens de programação de computadores



PLN

- Processamento de Língua Natural
 - Linguística Computacional
 - Processamento de Linguagem Natural
 - Engenharia das Línguas Naturais
- No Brasil, tradicionalmente visto como subárea da Inteligência Artificial & Computação
 - Habilidade linguística é um tipo de inteligência



PLN

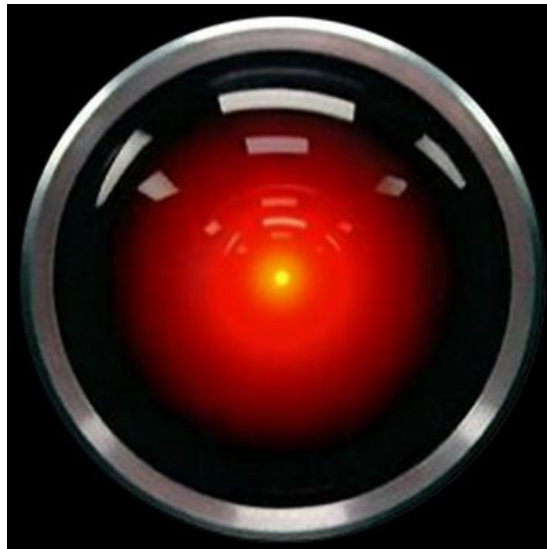
- Instruir o computador a lidar com a língua
 - Entendimento
 - Análise morfológica e sintática, semântica e discursiva
 - Geração, síntese
 - Tradução, produção de resumos
 - Correção gramatical
 - Busca de respostas para perguntas
 - Recuperação de informação da Internet
 - Auxílio a escrita e ao aprendizado de línguas
- Multidisciplinar
 - Computação
 - Linguística



META CLÁSSICA DA IA E DO PLN

- **HAL 9.000** (Heuristically programmed **AL**gorithmic Computer)
 - Incrível capacidade de linguagem
 - Inspiração clássica da IA e do PLN

1968



STANLEY KUBRICK'S
2001:
a space odyssey

CONVERSA CLÁSSICA COM HAL

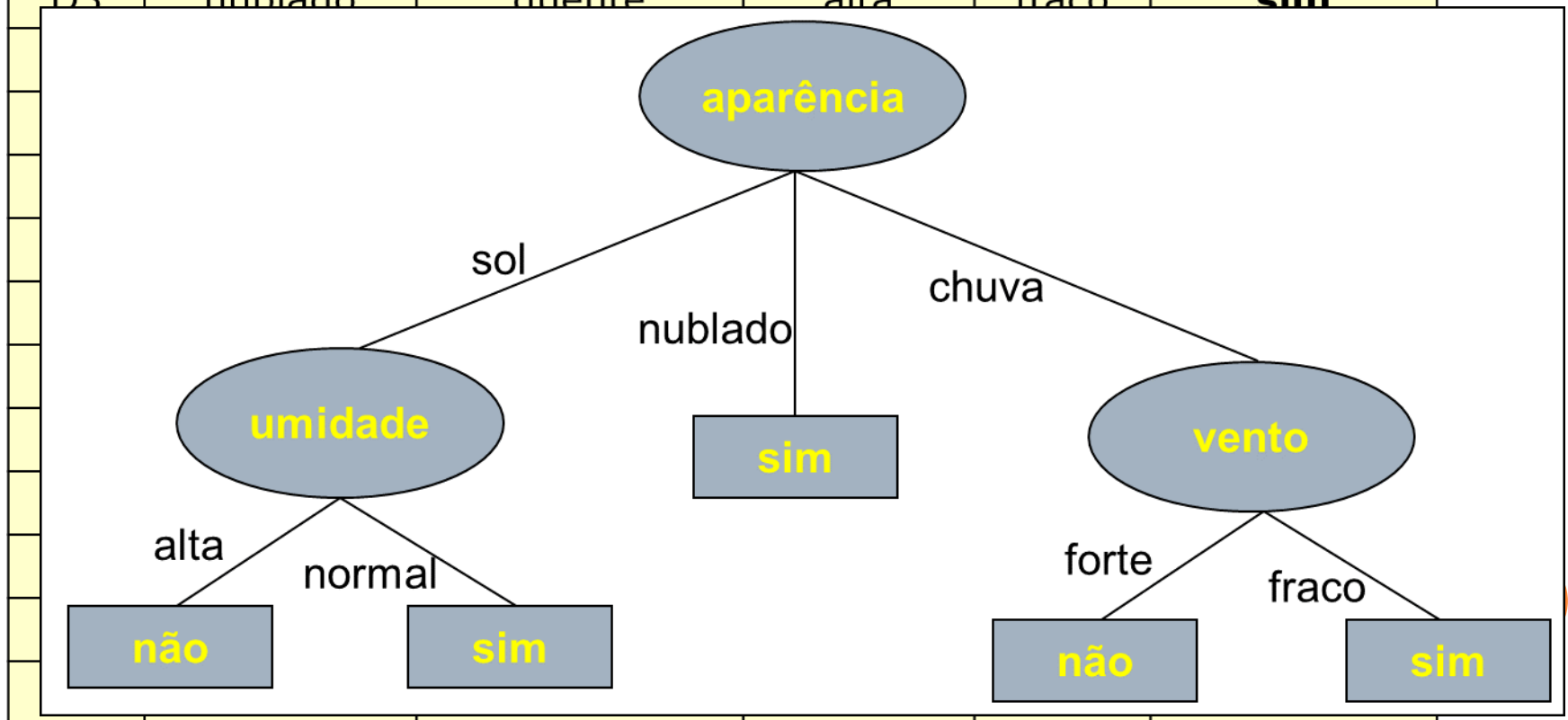


IA & PLN: UM POUCO DA HISTÓRIA




- Origem
- Expectativas e frustrações
- Renascimento
 - Inteligência
 - Redimensionamento
- Sistemas inteligentes
- Aprendizado de máquina
- Processamento de linguagem natural
 - De abordagens simbólicas a aprendizado automático
- *Big data*
- Ciência de dados

O QUE SE PODE APRENDER?

dia	aparência	temperatura	umidade	vento	jogar_tênis
D1	ensolarado	quente	alta	fraco	não
D2	ensolarado	quente	alta	forte	não
D3	nublado	quente	alta	fraco	sim

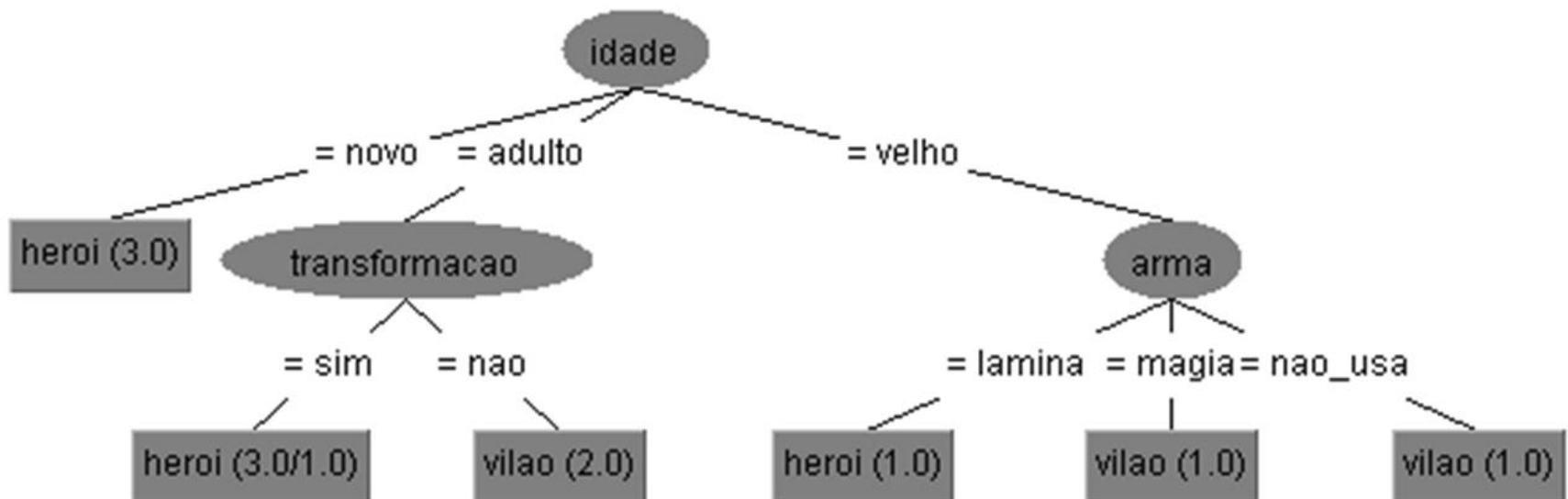



O QUE SE PODE APRENDER?

Personagem	Arma	Transformação	Idade	Classe
 He-man	Lâmina	Sim	Adulto	Herói
 Seiya	Magia	Não	Novo	Herói
 Mun-ra	Magia	Sim	Velho	Vilão

Weka Classifier Tree Visualizer: 15:46:14 - trees.J48 (herois)

Tree View



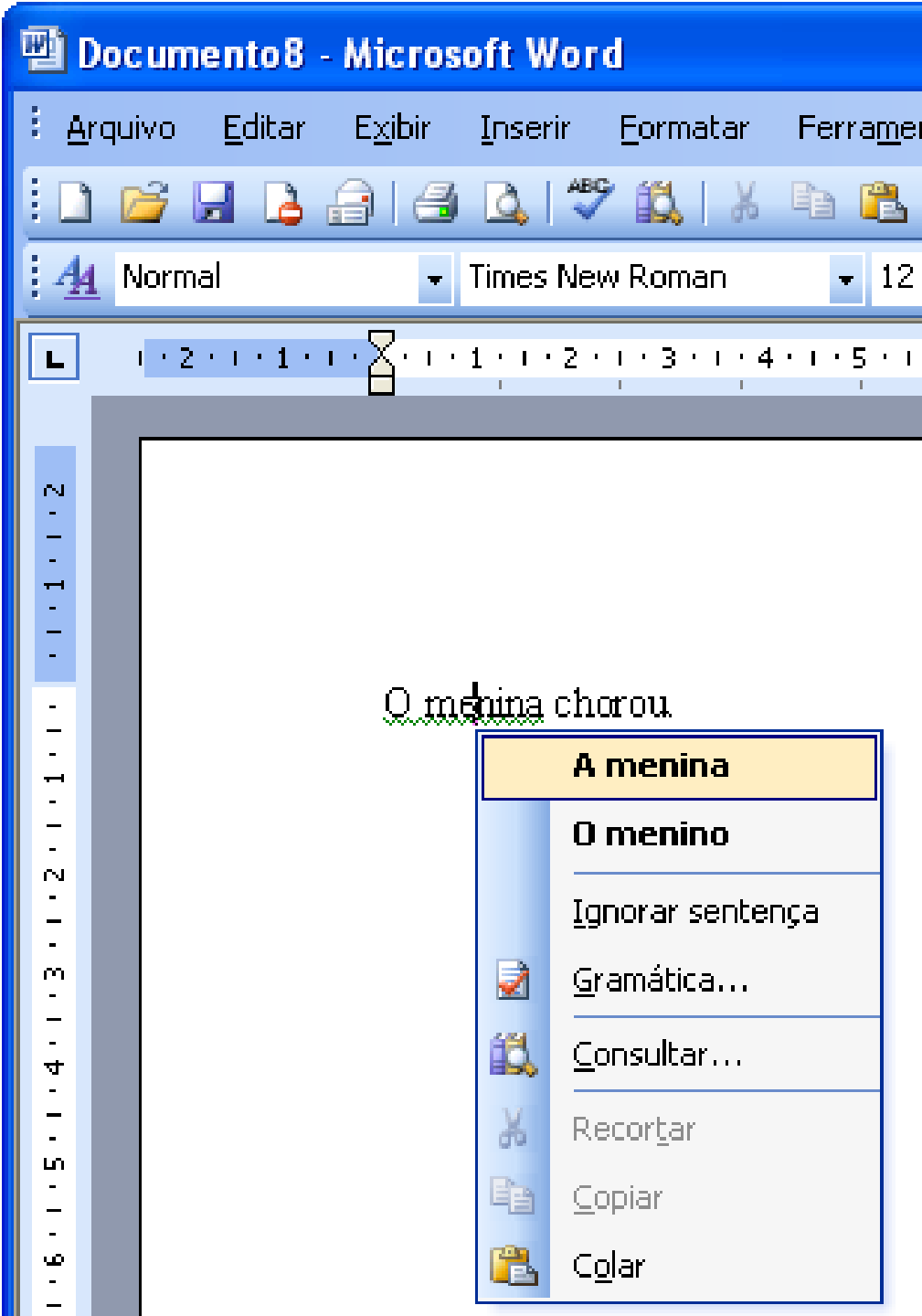
	Ben 10	Nao usa	Sim	Novo	Herói
---	--------	---------	-----	------	-------



**ALGUNS
EXEMPLOS
INSPIRADORES EM
PLN**

REVISÃO GRAMATICAL

- Basicamente, *análise sintática* e regras de correção gramatical



SPAMS

- Filtro de Spams: **para aquecer!**

Subject: PLEASE READ CAREFUL AND ACT AS INSTRUCTED

Dear Beneficiary

This is to officially inform you that we have written to you before without getting respond from you and we believe that our previous mail did not get to you therefore we write you again. We are contacting you concerning the release of your inheritance fund / Draft /Cheque /ATM Card which have been delayed for transfer by some officials who claim to be in position of your fund thereby extorting money from you in one way or the other.

Your Fund has finally been approved for transfer by the West Africa Fund Monitoring Unit. Your fund will be transfer to you via MasterCard ATM which is cash able in any ATM machine or Bank anywhere in the world.

...

Funciona bem?

SUMARIZAÇÃO

- Produção automática de resumo pela seleção de sentenças do texto-fonte

Folha de São Paulo

Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo. Segundo uma porta-voz da ONU, o avião, de fabricação russa, estava tentando aterrissar no aeroporto de Bukavu em meio a uma tempestade. **A aeronave se chocou com uma montanha e caiu, em chamas, sobre uma floresta a 15 quilômetros de distância da pista do aeroporto.** Acidentes aéreos são freqüentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética. O avião acidentado, operado pela Air Traset, levava 14 passageiros e três tripulantes. **Ele havia saído da cidade mineira de Lugushwa em direção a Bukavu, numa distância de 130 quilômetros.** Aviões são usados extensivamente para transporte na República Democrática do Congo, um vasto país no qual há poucas estradas pavimentadas. Em março, a União Européia proibiu quase todas as companhias aéreas do Congo de operar na Europa. Apenas uma manteve a permissão. Em junho, a Associação Internacional de Transporte Aéreo incluiu o Congo num grupo de vários países africanos que classificou como uma vergonha para o setor.

SUMARIZAÇÃO

- Kupiec et al. (1995)
 - Atributos avaliados
 - Tamanho mínimo de sentença (5 palavras)
 - Presença de palavras indicativas (por exemplo, “*in conclusion*”)
 - Posição da sentença no parágrafo (início, meio ou fim)
 - Presença de palavras-chave
 - Presença de palavras capitalizadas
 - Decisão: “vai para o sumário” (sim ou não)
 - “Sim” para as sentenças que aparecem em sumários humanos
 - Acurácia de 84%

ESCRITA BONITA (*GREAT WRITING*)

- Textos lindamente escritos, informativos e divertidos

David Quammen (Harper's magazine)

One morning early last winter a small item appeared in my local newspaper announcing the birth of an extraordinary animal. A team of researchers at Texas A&M University had succeeded in cloning a whitetail deer. Never done before. The fawn, known as Dewey, was developing normally and seemed to be healthy. He had no mother, just a surrogate who had carried his fetus to term. He had no father, just a "donor" of all his chromosomes. He was the genetic duplicate of a certain trophy buck out of south Texas whose skin cells had been cultured in a laboratory. One of those cells furnished a nucleus that, transplanted and rejiggered, became the DNA core of an egg cell, which became an embryo, which in time became Dewey. So he was wildlife, in a sense, and in another sense elaborately synthetic. This is the sort of news, quirky but epochal, that can cause a person with a mouthful of toast to pause and marvel. What a dumb idea, I marveled.

ESCRITA BONITA

- Louis e Nenkova (2013)
 - Atributos principais avaliados
 - Número total de palavras visuais
 - Número de palavras visuais em partes do artigo (início, meio e fim)
 - Número de imagens
 - Número de referências a pessoas
 - Número de referências a entidades inanimadas
 - Número de orações relativas
 - Número de palavras não usuais
 - Número de sequências fonéticas não usuais
 - Número de sequências não usuais de palavras
 - Proporção de sentenças narrativas, atributivas e de entrevistas
 - Proporção de palavras afetivas
 - Proporção de palavras relacionadas a pesquisa
 - Decisão: texto “típico” ou “bonito” (*great*)
 - Acurácia acima de 75%

ESCRITA BONITA

○ Louis e Nenkova (2013)

- Palavras visuais: *grass, mountain, green, hill, blue, field, brown, sand, desert, dirt, landscape, sky*
- Palavras não usuais: *undersheriff, woggle, ahmok, hofman, volga, oceanaut, trachoma, baneful, truffler, acriminal, corvair, entomopter*
- Sequências fonéticas não usuais: *showroom, yahoo, dossier, powwow, plowshare, oomph, chihuahua, ionosphere, boudoir, superb, zaire, oeuvre*
- Pares de palavras não usuais: *plasticky woman, psychogenic problems, yoplait television, physiologically do, subminimal level, amuck run, ehatchery investment, illegitimately put*

DETECÇÃO DE FAKE NEWS

○ Como diferenciar casos verdadeiros de falsos?

Fake	True
<p><i>Michel Temer propõe fim do carnaval por 20 anos, “PEC dos gastos”. Michel Temer afirmou que não deve haver gastos com aparatos supérfluos sem pensar primeiramente na educação do Brasil. A medida pretende cancelar o carnaval de 2018.</i></p>	<p><i>Michel Temer não quer o fim do Carnaval por 20 anos. Notícias falsas misturam proximidade dos festejos, crise econômica e medidas impopulares do governo do peemedebista.</i></p>
<p><i>Acabou a mordomia ! Ingresso mais barato pra mulher é ilegal. Baladas que davam meia entrada para mulher, ou até mesmo gratuidade, esto na ilegalidade agora. Acabou o preconceito com os homens nas casas de show de todo o Brasil.</i></p>	<p><i>Ingresso feminino barato como marketing ‘não inferioriza mulher’, diz juíza do DF. Afirmação consta em decisão sobre preços diferentes para homens e mulheres em festa no Lago Paranoá. ‘Prática permite que mulher possa optar por participar de tais eventos sociais’, diz texto.</i></p>

ESTATÍSTICAS DE UM CÓRPUS PARA O PORTUGUÊS (MONTEIRO ET AL., 2018)

Category	Number of samples	%
Politics	4,180	58.0
TV & celebrities	1,544	21.4
Society & daily news	1,276	17.7
Science & technology	112	1.5
Economy	44	0.7
Religion	44	0.7

ESTADÍSTICAS

Traditional features	Fake News	True News
Avg number of tokens	216.1	1,268.5
Avg number of types (without punctuation and numbers)	119.2	494.1
Avg size of words (in characters)	4.8	4.8
Type-token ratio	0.68	0.47
Avg number of sentences	12.7	54.8
Avg size of sentences (in words)	15.3	21.1
Avg number of verbs (norm. by the avg number of tokens)	14.3	13.4
Avg number of nouns (norm. by the avg number of tokens)	24.5	24.6
Avg number of adjectives (norm. by the avg number of tokens)	4.1	4.4
Avg number of adverbs (norm. by the avg number of tokens)	3.7	4.0
Avg number of pronouns (norm. by the avg number of tokens)	5.0	5.2
Avg number of stopwords (norm. by the avg number of tokens)	31.0	32.8
Percentage of news with spelling errors	36.0	3.0

ESTATÍSTICAS

Features Zhou et al. (2004)	Fake News	True News
Avg pausality per text (proportion of pauses)	2.46	3.04
Avg emotiveness per text (proportion of adjectives and adverbs)	0.20	0.21
Avg uncertainty per text (proportion of modal verbs and passive voice)	4.48	23.24
Avg non-immediacy per text (proportion of 1st and 2nd pronouns)	0.62	4.05

RESULTADOS

- Já em 97% de acerto
 - Muito a avançar: meias verdades, pós-verdades, etc.
- Website e aplicativo no WhatsApp
 - <http://nilc-fakenews.herokuapp.com>

PARA TRABALHAR COM PLN

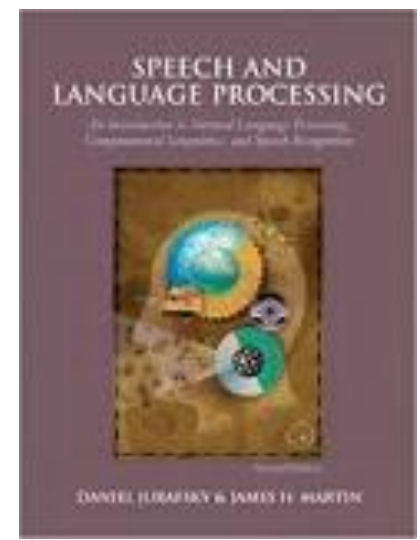
- Base: optativa do 6º período
 - SCC0230 Inteligência Artificial
- Disciplinas específicas
 - BCC: SCC0633 Processamento de Linguagem Natural (períodos ímpares)
 - BSI: SCC0532 Tópicos Avançados em Inteligência Artificial (próximo semestre)
- Gosto por
 - Matemática, estatística e computação
 - Línguas
 - Desafios!

Grupo de pesquisa: **NILC**

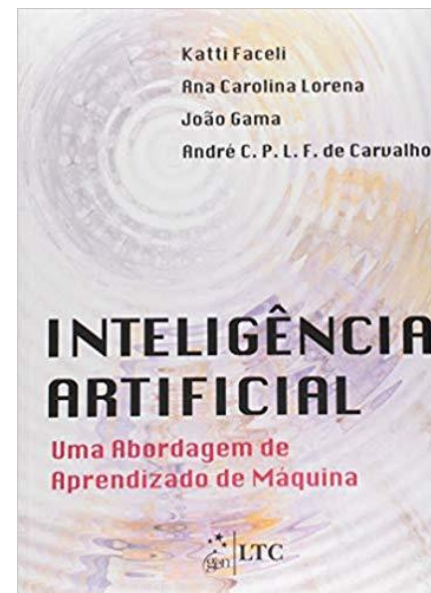
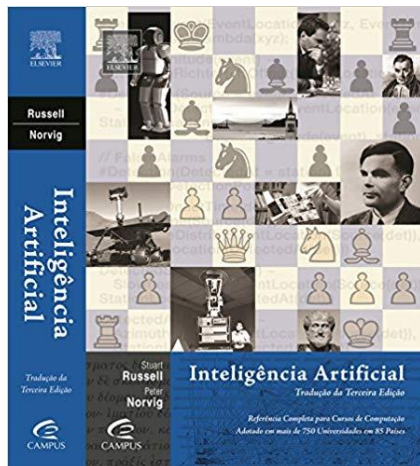
www.nilc.icmc.usp.br

MATERIAL COMPLEMENTAR PARA OS INTERESSADOS

- Livro já clássico de PLN
 - <https://web.stanford.edu/~jurafsky/slp3/>



- Livro clássico de IA
 - Russel e Norvig (última ed. de 2013)



- Livro mais recente do Prof. André Carvalho