

Processamento de Linguagem Natural em Tribunais de Contas

Alexandre Manir Figueiredo Sarquis

NLP em TC

- 1 Pré-processamento de texto para NLP
- 2 Representação de texto para NLP
- 3 "Tarefas" conhecidas e estudadas em NLP
- 4 Proposta de "tarefas" e pipelines adaptadas ao âmbito dos Tribunais de Contas
- 5 Grupos de pesquisa e fontes de modelos
- 6 Modelos não específicos para Tribunal de Contas (acompanhando no python - colab):
- 7 Exemplo de Pipeline completa aplicada a um caso do TC (se der tempo, código no GIT)

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

INTRODUÇÃO

- | | |
|---------------------------|---------------------|
| 1. Fontes | Controlabilidade |
| 2. Modelo | 6. Tensor |
| 3. Geometria
Analítica | 7. O que é um robô? |
| 4. Álgebra Linear | 8. Cognição humana |
| 5. Observabilidade e | 9. Nosso objetivo |

INTRODUÇÃO

O campo atualmente conhecido como “inteligência artificial” é uma complexa combinação de áreas.

A abordagem que analisaremos aqui, se pode ser definida, será a álgebra linear computacional.

Elaine Rich. Inteligência Artificial. McGraw-Hill, 1988.

George Casella & Roger Berger. Statistical Inference. Duxbury, 2002.

Gilbert Strang. Linear Algebra and Learning from Data. Wellesley-Cambridge, 2019.

Elon Lages Lima. Geometria Analítica e Álgebra Linear. IMPA, 2005.

Jan Goyvaerts & Steven Levithan. Regular Expressions Cookbook. O ‘Reilly, 2012.

Ashish Vaswani et ali. Attention is All You Need. Cornell, 2017.
[\(https://arxiv.org/abs/1706.03762\)](https://arxiv.org/abs/1706.03762)

INTRODUÇÃO

1. Sandra Kublik, Shubham Saboo. **GPT-3 Building Innovative NLP Products Using Large Language Models.** O'reilly, 2022.
2. Lewis Tunstall, Leandro von Werra e Thomas Wolf. **Natural Language Processing with Transformers.** O'Reilly, 2022.
3. Shashank Mohan Jain. **Introduction to Transformers for NLP.** Apress, 2022.
4. Thushan Ganegedara. **Natural Language Processing with TensorFlow.** Packt, 2022.
5. Akshay Kulkarni e Adarsha Shivananda. **Natural Language Processing Recipes.** Apress, 2021.
6. Aman Kedia e Mayank Rasu. **Hands-On Python Natural Language Processing.** BPB publications, 2021.
7. Ankur A. Patel e Ajay Uppili Arasanipalai. **Applied Natural Language Processing in the Enterprise.** O'Reilly, 2021.
8. Amandeep. **Implement NLP use-cases using BERT.** Bpb, 2021.
9. Duygu Altinok. **Mastering spaCy.** Packt, 2021.
10. Savaş Yıldırım, Meysam Asgari-Chenaghlu. **Mastering Transformers.** Packt, 2021.
11. Steve Tingiris. **Exploring GPT-3.** Packt, 2021.
12. Sudharsan Ravichandiran. **Getting Started with Google BERT.** Packt, 2021.
13. Sunil Patel. **Getting started with Deep Learning for Natural Language Processing.** BPB, 2021.
14. Zhenya Antić. **Python Natural Language Processing Cookbook.** Packt, 2021.
15. Butch Quinto. **Next-Generation Machine Learning with Spark.** Apress, 2020.
16. Mathangi Sri. **Practical Natural Language Processing with Python.** Apress, 2020.
17. Yuli Vasiliev. **Natural Language Processing wit.** No starch, 2020.
18. Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta e Harshit Surana. **Practical Natural Language Processing.** O'reilly, 2020.
19. Antonio Gulli, Amita Kapoor e Sujit Pal. **Deep Learning with TensorFlow 2.0 and Keras.** Packt, 2019.
20. Denis Rothman. **Transformers for Natural Language Processing.** Apress, 2019.
21. Hobson Lane, Hannes Hapke e Cole Howard. **Natural Language Processing in Action.** Manning, 2019.
22. Akshay Kulkarni e Adarsha Shivananda. **Natural Language Processing Recipes.** Apress, 2019.
23. Delip Rao e Brian McMahan. **Natural Language Processing with PyTorch.** O'reilly, 2019.
24. Uday Kamath, John Liu e James Whitaker. **Deep Learning for NLP and Speech Recognition.** Springer, 2019.
25. Benjamin Bengfort, Tony Ojeda e Rebecca Bilbro. **Applied Text Analysis with Python.** O'Reilly, 2018.
26. Dipanjan Sarkar, Raghav Bali, Tushar Sharma. **Practical Machine Learning with Python.** Apress, 2018.
27. Li Deng e Yang Liu. **Deep Learning in Natural Language Processing.** Springer, 2018.
28. Bhargav Srinivasa-Desikan. **Natural Language Processing and Computational Linguistics.** Packt, 2018.
29. Palash Goyal, Sumit Pandey e Karan Jain. **Deep Learning for Natural Language Processing.** Apress, 2018.
30. Andreas C. Müller e Sarah Guido. **Introduction to Machine Learning with Python.** O'Reilly, 2016.
31. Deepti Chopra, Nisheeth Joshi e Iti Mathur. **Mastering Natural Language Processing with Python.** Packt, 2016.
32. Steven Bird, Ewan Klein e Edward Loper. **Natural Language Processing with Python.** O'Reilly, 2009.

INTRODUÇÃO

Precisamos então de:

- Matemática – geometria e álgebra linear
- Ciência da computação – bases de dados
- Ciência da computação – programação – python e alguns pacotes
- Ciência da computação – redes neurais
- Linguística – funções sintáticas, semântica (palavra tomada dentro da oração)
- Linguística – funções léxicas, morfológica, grammatical (palavra tomada isoladamente)
- Área fim – contratos, licitações, processo de contas

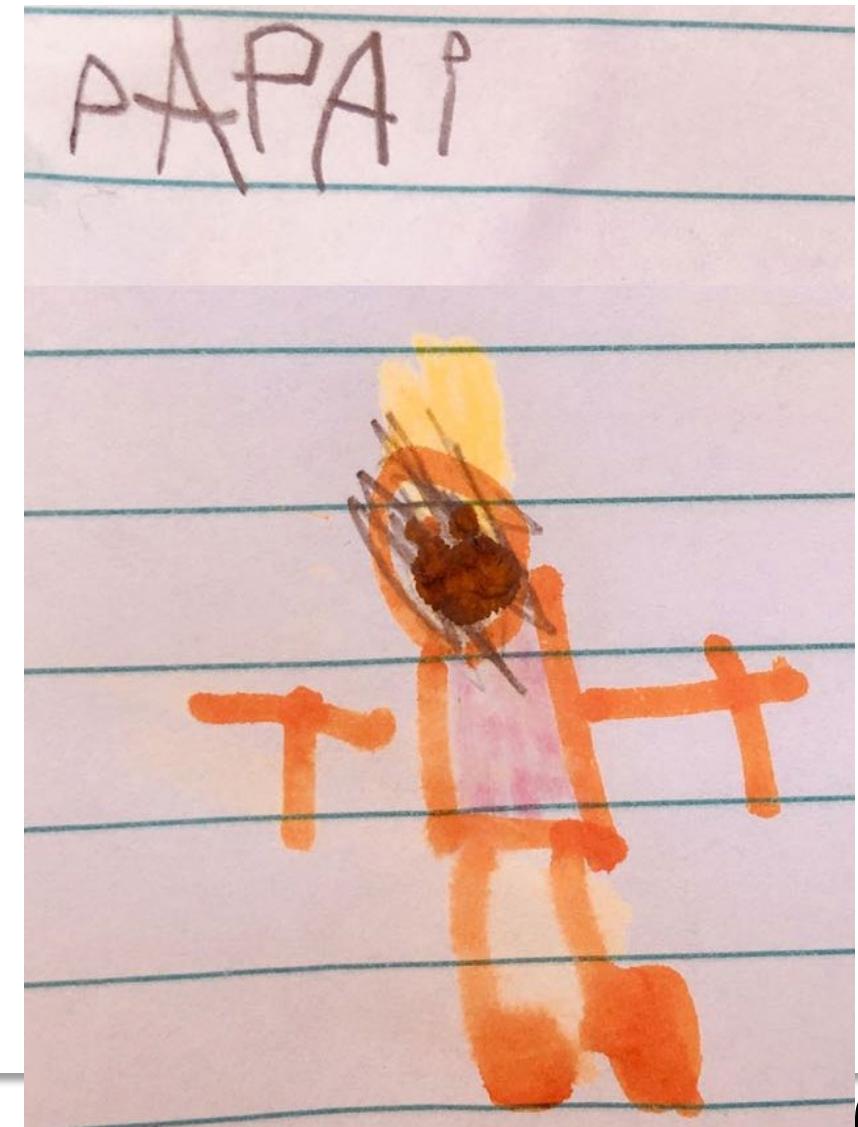
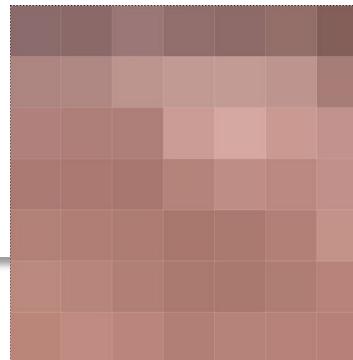
Mas será que precisamos conhecer todos a fundo?

INTRODUÇÃO



Favor não entrar em pânico nos próximos slides, pois precisamos de apenas alguns dos conceitos!

INTRODUÇÃO



INTRODUÇÃO

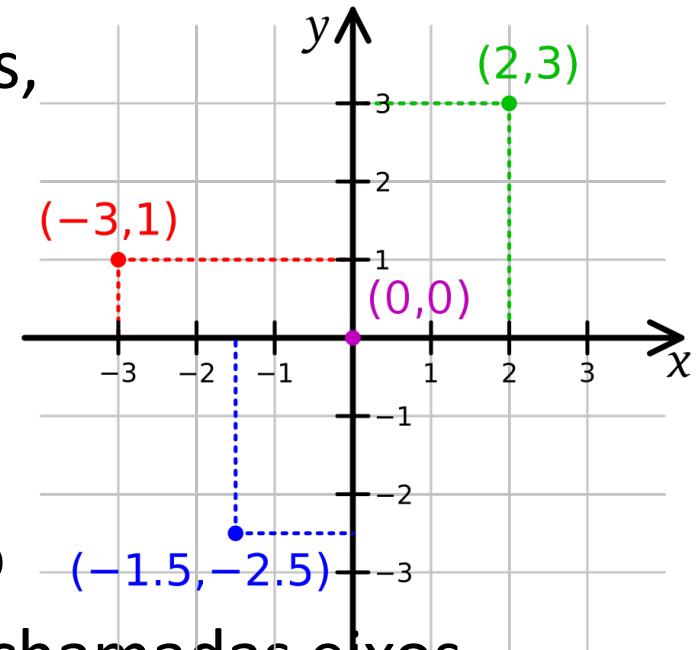
Geometria Analítica

Área da geometria muito ligada ao sistema de coordenadas cartesianas.

Geometria é a área da matemática que estuda o espaço, para tanto analisando estruturas como distâncias, pontos, linhas, curvas, ângulos e superfícies. Geometria analítica é o estudo da geometria no âmbito de um sistema coordenado.

Sistema de coordenadas cartesianas é aquele em que cada ponto é representado unicamente por um conjunto de distâncias do ponto a linhas perpendiculares entre si chamadas eixos.

https://en.wikipedia.org/wiki/Analytic_geometry#/media/File:Cartesian-coordinate-system.svg



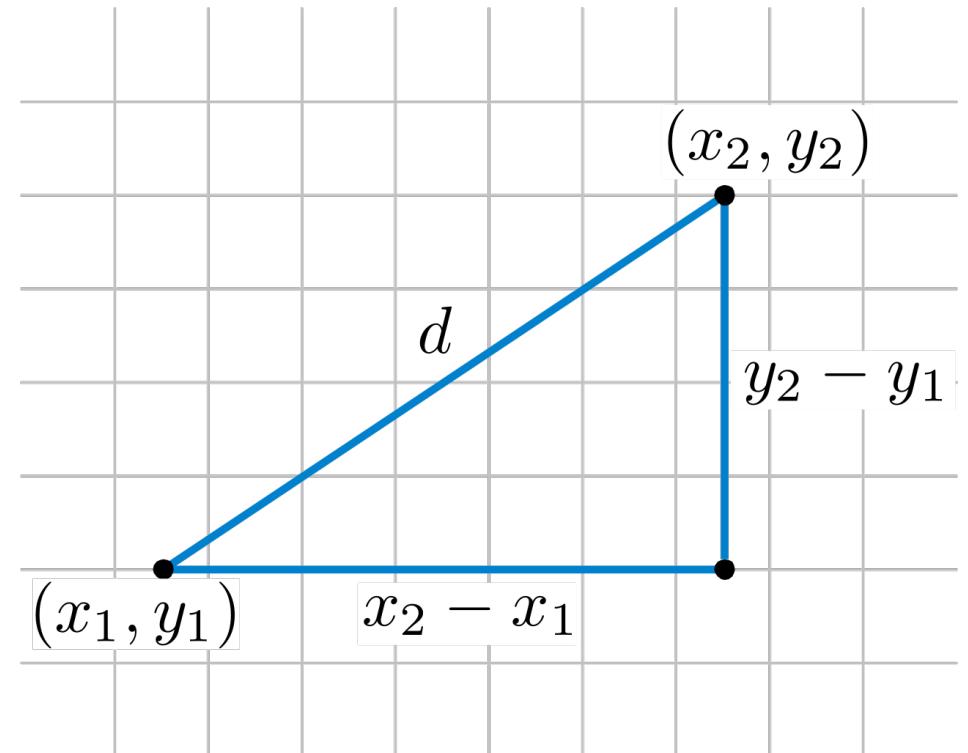
INTRODUÇÃO

Geometria Analítica – “distância”

Teorema de Pitágoras

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

https://en.wikipedia.org/wiki/Analytic_geometry#/media/File:Distance_Formula.svg



INTRODUÇÃO

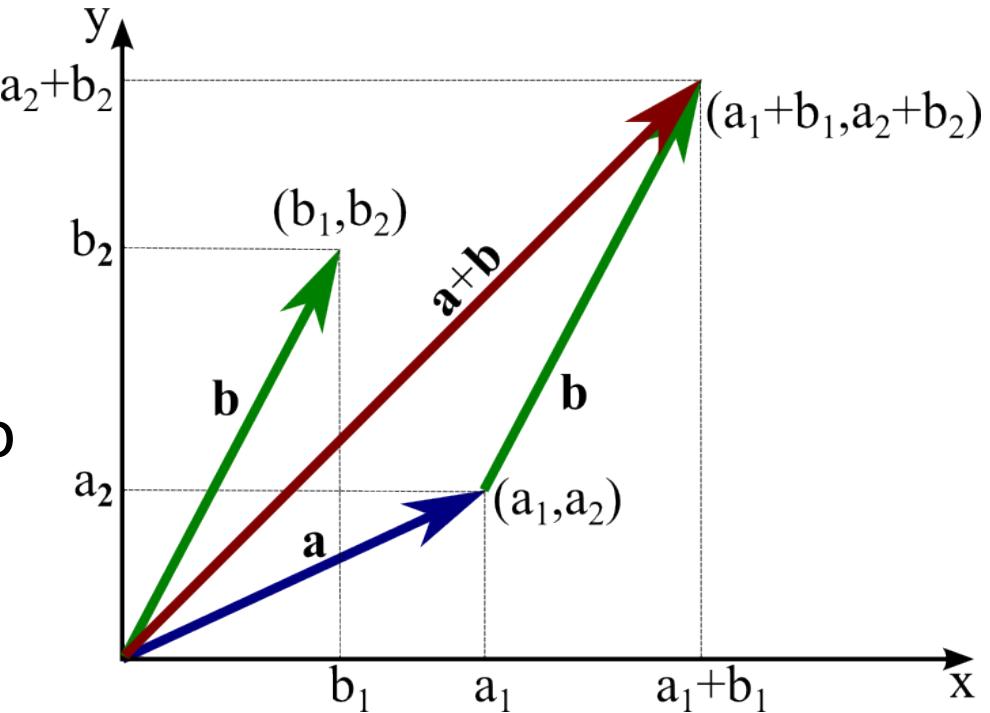
Geometria Analítica – “vetor”

Objeto que não pode ser representado por um único número?

Objeto matemático que representa a translação de um ponto?

Quantas coordenadas possui um vetor?

$$\sum_{j=0}^{N-1} s_j \cdot e^{-i \cdot 2\pi \cdot \frac{k \cdot j}{N}}$$



https://mathinsight.org/vectors_cartesian_coordinates_2d_3d

INTRODUÇÃO

Geometria Analítica

Cada coordenada de um vetor pode ser, portanto, uma medição de um atributo do objeto sendo estudado que é absolutamente independente das demais medidas.

Se ocorre a dependência, a base coordenada não é “ortogonal”.

Ademais, é útil que a medida represente a exata quantidade daquele atributo (metros, unidades, segundos, etc).

Se a medida não é unitária, a base coordenada não é “normal”.

Nosso ideal é, portanto, uma base coordenada cartesiana ortonormal.

INTRODUÇÃO

Geometria Analítica – Controle Linear

Uma vez que generalizamos o conceito de coordenada, podemos chamar cada elemento de “estado”.

Se o conhecimento do objeto é completo, todos os estados são observáveis, pois são mensuráveis.

Na realidade, entretanto, há estados não diretamente mensuráveis que podem, entretanto e ainda assim, serem observáveis, se dos demais estados possam ser inferidos.

Há, também, estados que não podem ser mesmo observáveis.

INTRODUÇÃO

Geometria Analítica – Controle Linear – Estados não observáveis

Carga semântica é vetor? Carga léxica é vetor?

Significado x significante.

Figuras de linguagem – ironia, hipérbole, elipse.

Recursos gráficos - Itálico, negrito, taxado, nota de rodapé, parênteses.

Polissemia.

“Modos” verbais – indicativo, subjetivo e imperativo.

Técnicas narrativas – voz passiva, voz ativa e ocultação do sujeito.

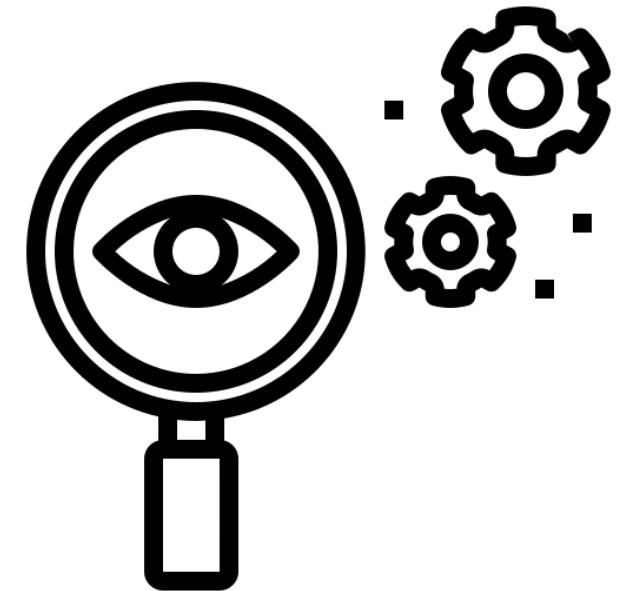
INTRODUÇÃO

Geometria Analítica – Controle Linear

“Controle” é o estímulo externo apto a ser aplicado a um sistema.

Se a partir de um dado estado do sistema for possível alcançar outro estado por meio de um controle, o estado é chamado controlável.

Estados não observáveis não são controláveis, mas mesmo estados observáveis podem não ser controláveis.



INTRODUÇÃO

Geometria Analítica – Distância

$$\|(z_1, z_2)\|_2 = \sqrt{z_1^2 + z_2^2}$$

O vetor que une em linha reta um ponto \vec{x} a um ponto \vec{y} é representado pela diferença

$$\vec{z} = \vec{x} - \vec{y}$$

O “tamanho” da linha reta que une os pontos deve ser a distância.

$$\|(z_1, z_2)\|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

INTRODUÇÃO

Geometria Analítica – Produto Interno

$$\langle \vec{x}, \vec{y} \rangle = x_1 \cdot y_1 + x_2 \cdot y_2$$

Projeção ortogonal

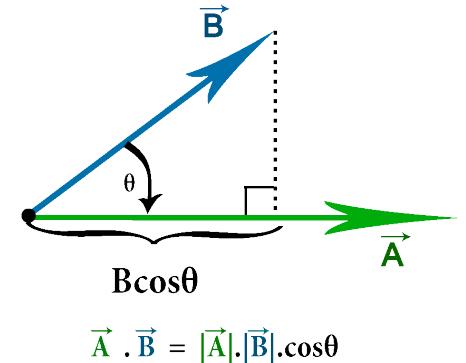
$$\langle \vec{a}, \vec{b} \rangle = a_1 \cdot b_1 + a_2 \cdot b_2 = \|\vec{a}\|_2 \cdot \|\vec{b}\|_2 \cdot \cos\theta$$

Tamanho

$$\langle \vec{a}, \vec{a} \rangle = a_1 \cdot a_1 + a_2 \cdot a_2 = \|\vec{a}\|_2^2$$

Distância

$$\langle \vec{a} - \vec{b}, \vec{a} - \vec{b} \rangle = (a_1 - b_1)^2 + (a_2 - b_2)^2 = \|D_{a,b}\|_2^2$$



INTRODUÇÃO

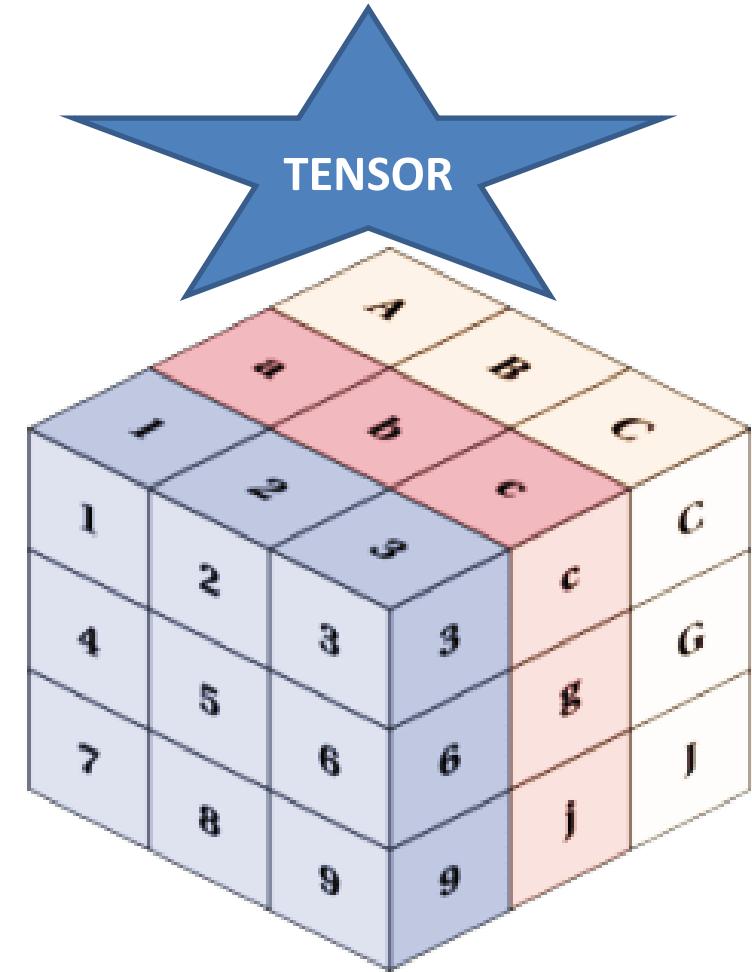
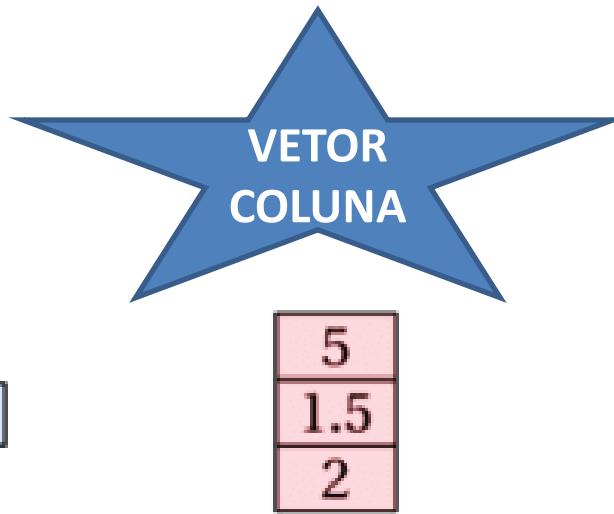
Geometria Analítica – Tensor

Se vetor é um conjunto de elementos, genericamente pensado como a generalização de um número, cada um dos elementos pode ser pensando, ele mesmo como um vetor, em uma generalização de vetor.

Se todos esses vetores forem vetores de números e de mesma dimensão, teremos uma matriz.

Se os vetores dos vetores forem eles mesmos vetores, irrelevante a coincidência de dimensões (embora geralmente sejam representados ainda com alguma coincidência) são chamados tensores.

INTRODUÇÃO



INTRODUÇÃO

Brevíssimas observações a respeito de regressão

Se o característica observada é uma categoria, é natural desejar espremer sua representação em apenas uma coordenada, por exemplo a categoria morfológica (0 =substantivo, 1=verbo, 2=advérbio, 3=pronome, 4=adjetivo, 5=artigo, etc.).

Conferida à análise numérica, entretanto, essa representação é problemática:
Substantivo (0) representa a ausência de verbo (1)?

A presença de um adjetivo (4) é duas vezes mais intensa para o fenômeno medido que a presença de um advérbio (2)?

Um verbo (1) e um pronome (3) podem ser substituídos por um adjetivo (4)?

Em inexistindo uma sexta categoria (6) pode alguns de seus desdobramentos ser inferidos a partir das demais categorias, por extração?

INTRODUÇÃO

Brevíssimas observações a respeito de regressão

“one-hot” encoding.

Se vamos presumir absolutamente nada a respeito do ângulo formado entre os vetores, o mais seguro é representá-los ortogonais.

	1	2	3	4	5	6	7
eu	1	0	0	0	0	0	0
comi	0	1	0	0	0	0	0
uma	0	0	1	0	0	0	0
maça	0	0	0	1	0	0	0
e	0	0	0	0	1	0	0
toquei	0	0	0	0	0	1	0
piano	0	0	0	0	0	0	1

INTRODUÇÃO

Brevíssimas observações a respeito de regressão

Multicolinearidade

Endogeneidade

Não estacionariedade

	1	2	3	4	5	6	7
eu	1	0	0	0	0	0	0
comi	0	1	0	0	0	0	0
uma	0	0	1	0	0	0	0
maça	0	0	0	1	0	0	0
e	0	0	0	0	1	0	0
toquei	0	0	0	0	0	1	0
piano	0	0	0	0	0	0	1



INTRODUÇÃO

Vamos ficar um pouco filosóficos....

O que é robô?

Karel Čapek – Rossumovi Univerzální
Roboti

“roboť” 1920 – trabalho forçado

“Andróide” – robô humanóide



INTRODUÇÃO

Vamos ficar um pouco filosóficos....

Para o conceito de robô, é necessária a forma humanóide?

É necessária tarefa típica de um humano?

É necessário haver vontade e, nessa vontade, haver contrariedade?

Um rio não é um robô, mas um equipamento que recupera agua fresca, pode ser.

Parece que o tema central é a conversão de energia em utilidade de uma forma não orgânica.

Produção, transformação e classificação de texto – robô.

INTRODUÇÃO

Alegoria da Caverna e o conceito de modelo



https://commons.wikimedia.org/wiki/File:An_Illustration_of_The_Allegory_of_the_Cave,_from_Plato%27s_Republic.jpg

INTRODUÇÃO

Alegoria da Caverna e o conceito de modelo



https://commons.wikimedia.org/wiki/File:An_Illustration_of_The_Allegory_of_the_Cave,_from_Plato%27s_Republic.jpg

INTRODUÇÃO

Alegoria da Caverna e o conceito de modelo



<https://www.aerotime.aero/articles/31470-defects-found-in-components-of-cfm-leap-1a-engines-faa>

INTRODUÇÃO

Cognição Humana – conhecimento a priori e a posteriori

Conhecimento a priori – inato e congênito, não se trata de produto da experiência mas somente pode ser explicado em virtude de já estar presente ao nascimento, desnecessário qualquer treinamento.

Conhecimento a posteriori – decorre do que os sentidos propiciam ao intelecto conhecer.



https://snl.no/Immanuel_Kant

INTRODUÇÃO

Cognição Humana – conhecimento conotativo e denotativo

Conhecimento conotativo:

Cão – totó, rex e dino são cães.

Recorre à indução, à realidade e à extração da experiência.

É intrinsecamente controverso e equívoco. Um lobo é um cão?

Conhecimento denotativo:

Cão – animal doméstico que faz “au-au”.

Recorre à dedução, à redução, à síntese e à lógica.

É intrinsecamente incontroverso: alguns papagaios são cães.

INTRODUÇÃO

O que precisamos, enfim

“tensor”

“distância”

“modelo”

“estado não observável”

Superar a necessidade de ser preciso e de compreender todas as etapas intermediárias.

Superar a necessidade de que os problemas sejam integralmente solucionados.

Aquilatar as ferramentas pelas suas utilidades.

INTRODUÇÃO

Desafio daqui para diante:

- 1 – representar texto como tensores.
- 2 – aproveitar o conceito de distância.
- 3 – conhecer as técnicas de *deep learning*.
- 4 – apresentar o que tem sido feito com essas técnicas.
- 5 – avaliar se o nosso nicho contém problemas que possam se beneficiar de técnicas similares ou adaptadas.

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

FERRAMENTAS PRÉ-PROCESSAMENTO

- 1. Notepad++
- 2. Excel
- 3. HTML
- 4. Chrome
- 5. Regex
- 6. Postgres
- 7. Python
- 8. PyCharm,
Spyder, VSCode
ou etc?

FERRAMENTAS PRÉ-PROCESSAMENTO

Hedit.exe

Iso (ansi)

UTF-8

ASCII

The screenshot shows a hex viewer application window titled 'D:/Downloads/contratos etcesp v1.csv — Hex View'. The interface includes a toolbar with 'Open...', 'Offset 0', 'Encoding ASCII', and 'Quit' buttons. The main area displays the binary data of the CSV file in two columns: hex values and ASCII characters. The ASCII column contains fields such as 'numeroProcesso;c', 'odDocumento;nome', 'Arquivo;codArqui', 'vo;caminho;conte', 'ntType;autuacao;', 'status;exercicio', 'ambito;classe;r', 'esultado;relator', 'valor;contratan', 'te;DOCcontratant', 'e;contratada;DOC', 'contratada;resum', 'o;;;;..21119891', '33;242028;096a09', '8-M-RRM-TC-00211', '1-989-13+.pdf.p7', 's;155112;/2013/5', '/2111989133/arqu', 'ivol52887.p7s;ap', 'lication/pdf;27', '/08/2013 11:03;A', 'RQUIVADO;2013;M;', 'Contrato (INICIA', 'NULL;rob', 'sonmarinho;34050', 'PREFEITURA MU', 'NICIPAL DE TAUBA', 'TE;4517600500010', 'FORTNORT DESEN', 'VOLVIMENTO AMBIE', 'NTAL E URBANO EI', 'RELI;90084600018'. The hex values are color-coded in green and blue.

FERRAMENTAS PRÉ-PROCESSAMENTO

ASCII – cada byte, uma representação.

ISO, UTF e outros – há um byte ou conjunto de bytes que sinalizam que os subsequentes devem ser interpretados como uma representação só.

FERRAMENTAS PRÉ-PROCESSAMENTO

Character Name	Char	Code	Decimal	Binary	Hex
Null	NUL	Ctrl @	0	00000000	00
Start of Heading	SOH	Ctrl A	1	00000001	01
Start of Text	STX	Ctrl B	2	00000010	02
End of Text	ETX	Ctrl C	3	00000011	03
End of Transmit	EOT	Ctrl D	4	00000100	04
Enquiry	ENQ	Ctrl E	5	00000101	05
Acknowledge	ACK	Ctrl F	6	00000110	06
Bell	BEL	Ctrl G	7	00000111	07
Back Space	BS	Ctrl H	8	00001000	08
Horizontal Tab	TAB	Ctrl I	9	00001001	09
Line Feed	LF	Ctrl J	10	00001010	0A
Vertical Tab	VT	Ctrl K	11	00001011	0B
Form Feed	FF	Ctrl L	12	00001100	0C
Carriage Return	CR	Ctrl M	13	00001101	0D
Shift Out	SO	Ctrl N	14	00001110	0E
Shift In	SI	Ctrl O	15	00001111	0F
Data Line Escape	DLE	Ctrl P	16	00010000	10

FERRAMENTAS PRÉ-PROCESSAMENTO

Byte não é o mesmo que bytecode,
não é o mesmo que Unicode, não é
o mesmo que caracter, não é o
mesmo que dígito.

FERRAMENTAS PRÉ-PROCESSAMENTO

Elementos usuais de arquivos texto são:

- Cabeçalho.
- Separador de campos.
- Indicador de fim de linha.
- Codificação.
- Separador de milhar.
- Delimitador de texto (', ", "", "")
- Formatação de datas.

FERRAMENTAS PRÉ-PROCESSAMENTO

Dia/mês/ano

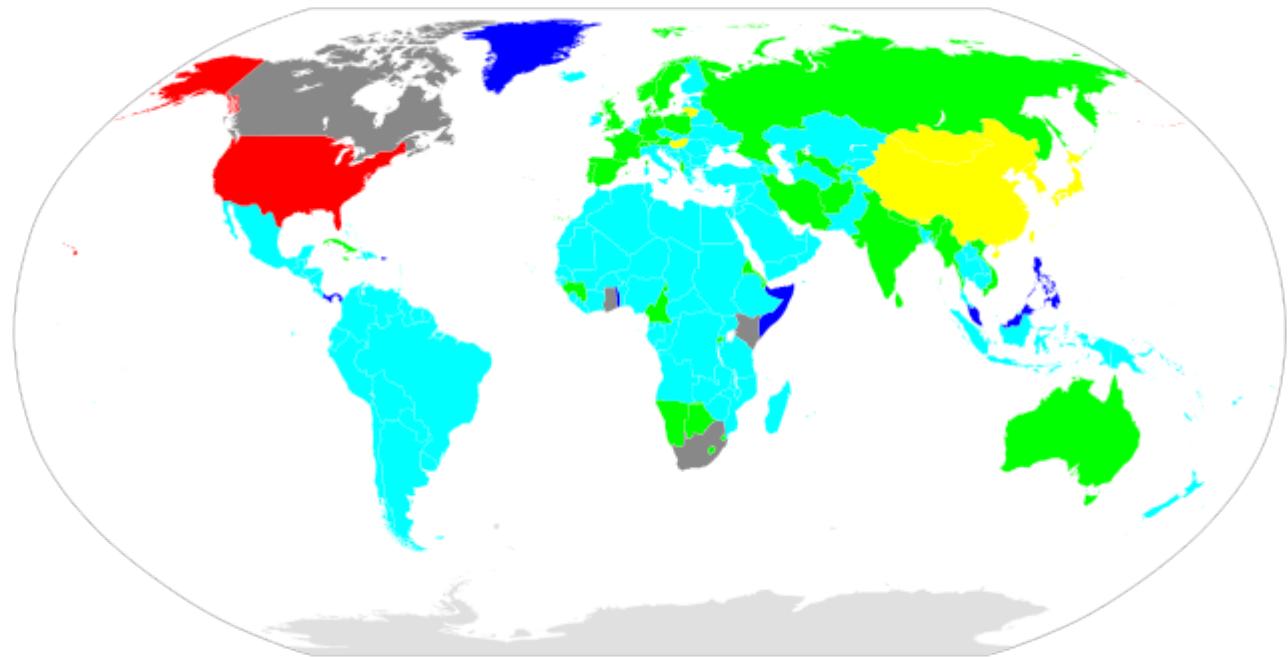
2865 mi usuários

Mês/dia/ano

328 mi usuários

Ano/mês/dia

2316 mi usuários



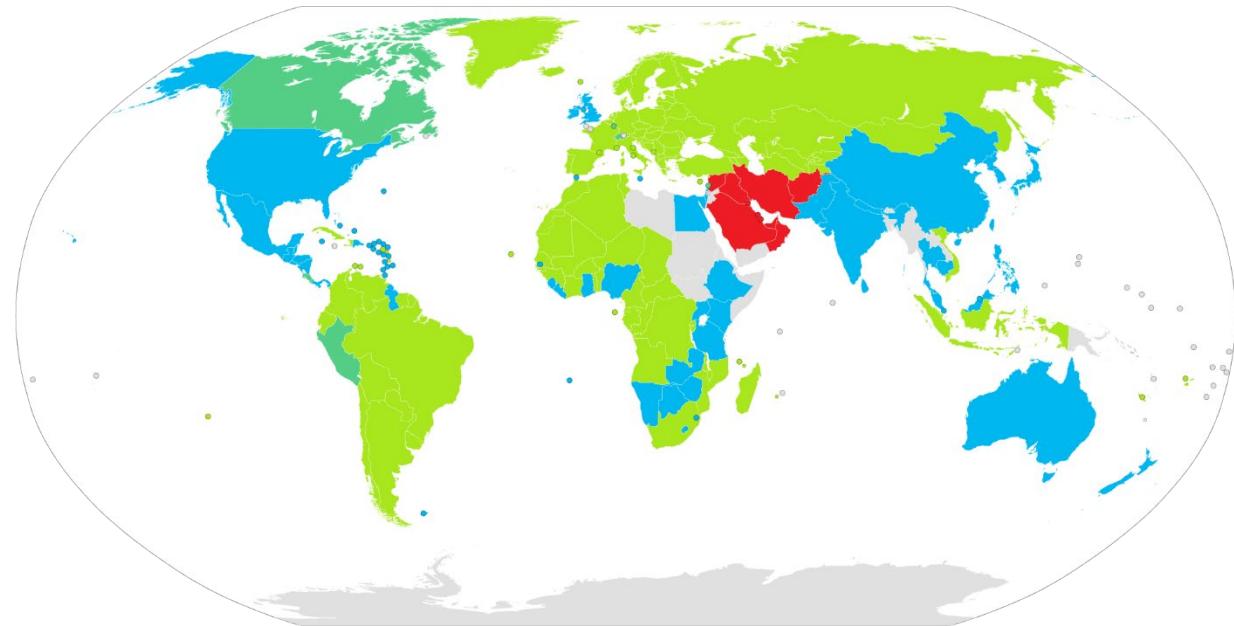
https://en.wikipedia.org/wiki/Date_format_by_country

FERRAMENTAS PRÉ-PROCESSAMENTO

Ponto

Vírgula

Ambos



https://en.wikipedia.org/wiki/Decimal_separator

FERRAMENTAS PRÉ-PROCESSAMENTO

Line Feed = 10 = #0A = LF

Ponto = 46 = #2E = “.”

Carriage Return = 13 = #0D = CR

Vírgula = 44 = #2C = “,”

Space = 32 = #20 = SPC

Ponto e vírgula = 59 = #3B = “;”

Tab = 9 = #09 = →

Escape do UTF-8 = 195 = #C3 = “”

Form Feed = 12 = #0C = FF

FERRAMENTAS PRÉ-PROCESSAMENTO

Final de linha “Windows”:

Um *carriage return* seguido de um *line feed*.

Final de linha “Unix”:

Apenas um *line feed*.

Final de linha “Apple”:

Apenas um *carriage return*.

No Word, o line feed sem carriage return é introduzido com shift+return e é representado pelo caracter ↵

FERRAMENTAS PRÉ-PROCESSAMENTO

D:\Downloads\contratos etcesp v1.csv - Notepad++

Arquivo Editar Localizar Visualizar Formatar Linguagem Configurações Ferramentas Macro Executar Plugins Janela ?

novo 21 novo 22 CARGA_USUARIOS.csv novo 23 novo 24 novo 25 novo 26 contratos etcesp v1.csv contrato.csv

```
1 numeroProcesso;codDocumento;nomeArquivo;codArquivo;caminho;contentType;autuacao;status;exercicio;ambito;classe;resultado;relator;valor;contr
2 2111989133;242028;096a098-M-RRM-TC-002111-989-13+.pdf.p7s;155112;/2013/5/2111989133/arquivol52887.p7s;application/pdf;27/08/2013.11:03;ARQUI
3 Concorrência Pública nº. 05-C/11. - CRLF
4 CONTRATO: s/n de 02/04/2013. - CRLF
5 OBJETO: Prestação de serviços de coleta de resíduos sólidos domiciliares em diversos logradouros públicos do município de Taubaté, com forne
6 VIGÊNCIA: 03/04/2013 a 02/10/2013.";;;;CRLF
7 3318989134;301543;TC-000120-989-13.pdf.p7s;193202;/2013/7/3318989134/arquivol90087.p7s;application/pdf;12/11/2013.17:13;ARQUIVADO;2013;E;Con
8 Pregão Eletrônico DRN nº. 001/2013. - CRLF
9 Processo RNE 0005/2013. - CRLF
10 Objeto a prestação de serviços de vigilância/segurança patrimonial CASA Franca, CASA Franca República e CAIP Franca. - CRLF
11 Vigência de 12 meses.";;;;CRLF
12 3351989132;301572;TC-000121-989-13.pdf.p7s;193224;/2013/7/3351989132/arquivol90109.p7s;application/pdf;13/11/2013.16:20;ARQUIVADO;2013;E;Con
13 Pregão Eletrônico DRN nº. 002/2013. - CRLF
14 Processo nº. 0006/2013. - CRLF
15 Objeto: a prestação de serviços de vigilância/segurança patrimonial para a Divisão Regional Norte, os Centros de Atendimento Socioeducativo.
16 A vigência do contrato é de 12 meses.";;;;CRLF
17 3407989136;301606;TC-000122-989-13.pdf.p7s;193237;/2013/7/3407989136/arquivol90122.p7s;application/pdf;18/11/2013.11:11;ARQUIVADO;2013;E;Con
18 2365989136;303751;041e042-M-RRM-TC-00002365-989-13-6+.pdf.p7s;194694;/2013/5/2365989136/arquivol91548.p7s;application/pdf;10/09/2013.16:14;A
19 ATA DE REGISTRO DE PREÇOS N°. 32/13 DE 25/04/13. - CRLF
20 Objeto: Registro de Preços para fornecimento de materiais escolares com manuseio para formação de kits e entrega ponto a ponto nas escolas d
21 1527989131;303758;047e048-M-RRM-TC-00001292-989-12-6+.pdf.p7s;194707;/2013/4/1527989131/arquivol91561.p7s;application/pdf;11/07/2013.16:40;A
22 Processo nº. 2012/56500. - CRLF
23 Objeto: Contratação de Empresa especializada na prestação de serviços de administração de cartão de combustível, através de cartão magnético
24 1847989134;309629;000633-989-13-e outros.pdf.p7s;198345;/2013/4/1847989134/arquivol95103.p7s;application/pdf;07/08/2013.15:20;ARQUIVADO;2013
25 OBJETO: Aquisição de kits de material escolar e mochila escolar a serem distribuídos aos alunos da rede municipal de ensino. - CRLF
26 VIGÊNCIA: 18/06/2013 a 17/06/2014";;;;;CRLF
27 1862989134;309658;000633-989-13-e outros.pdf.p7s;198360;/2013/4/1862989134/arquivol95118.p7s;application/pdf;08/08/2013.16:15;ARQUIVADO;2013
28 Objeto: Aquisição de kits de material escolar e mochila escolar a serem distribuídos aos alunos da Rede Municipal de Ensino
29 VIGÊNCIA: 18/06/2013 a 17/06/2013CRLF
30 ".....CRLF
```

gues;176869;PREFEITURA I

42 2A
43 2B
44 2C
45 2D
46 2F

12 Windows (CR LF) UTF-8

FERRAMENTAS PRÉ-PROCESSAMENTO

Espaços, hífens, parágrafos e outras pontuações podem ter diversas representações subjacentes.

Desejamos conduzir às mesmas coordenadas, sob pena de permitir o tratamento dos mesmos significados de forma desigual.

FERRAMENTAS PRÉ-PROCESSAMENTO

Code	Name of the character	Sample	Width of the character
U+0020	SPACE	foo bar	Depends on font, typically 1/4 em, often <u>adjusted</u>
U+00A0	NO-BREAK SPACE	foo bar	As a space, but often not adjusted
U+1680	OGHAM SPACE MARK	foo–bar	Unspecified; usually not really a space but a dash
U+180E	MONGOLIAN VOWEL SEPARATOR	foobar	0
U+2000	EN QUAD	foo bar	1 en (= 1/2 em)
U+2001	EM QUAD	foo bar	1 em (nominally, the height of the font)
U+2002	EN SPACE (nut)	foo bar	1 en (= 1/2 em)
U+2003	EM SPACE (mutton)	foo bar	1 em
U+2004	THREE-PER-EM SPACE (thick space)	foo bar	1/3 em
U+2005	FOUR-PER-EM SPACE (mid space)	foo bar	1/4 em
U+2006	SIX-PER-EM SPACE	foo bar	1/6 em
U+2007	FIGURE SPACE	foo bar	“Tabular width”, the width of digits
U+2008	PUNCTUATION SPACE	foo bar	The width of a period “.”
U+2009	THIN SPACE	foo bar	1/5 em (or sometimes 1/6 em)
U+200A	HAIR SPACE	foo bar	Narrower than THIN SPACE
U+200B	ZERO WIDTH SPACE	foobar	0
U+202F	NARROW NO-BREAK SPACE	foo bar	Narrower than NO-BREAK SPACE (or SPACE), “typically the width of a thin space or a mid space”
U+205F	MEDIUM MATHEMATICAL SPACE	foo bar	4/18 em
U+3000	IDEOGRAPHIC SPACE	foo bar	The width of ideographic (CJK) characters.
U+FEFF	ZERO WIDTH NO-BREAK SPACE	foobar	0

FERRAMENTAS PRÉ-PROCESSAMENTO

Exercício:

Nosso objetivo é deixar o arquivo
“contratos etcesp v1 poluído.csv”
No formato Windows, UTF-8, separador “;”
<https://github.com/alexandre-sarquis/NLP>

FERRAMENTAS PRÉ-PROCESSAMENTO

E o Excel?

Os arquivos texto “comma separated values” podem ser importados.

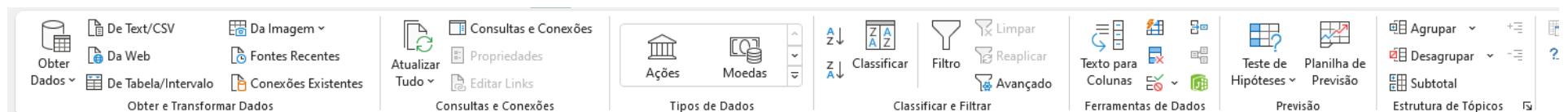
Se houver cópia de raw text, a formatação deve ser Ansi, o separador de coluna é o tab, o fim de linha é formato Windows.

FERRAMENTAS PRÉ-PROCESSAMENTO

A screenshot of a Microsoft Excel spreadsheet. The ribbon at the top shows the 'Dados' tab is selected. The 'Obter e Transformar Dados' section is open, displaying options like 'De Text/CSV', 'Da Imagem', 'Consultas e Conexões', and 'Tipos de Dados'. The 'Tipos de Dados' group contains icons for 'Ações' and 'Moedas'. The 'Classificar e Filtrar' group includes 'Z↓ A↓' (Classificar), 'Filtro', 'Limpar', 'Reaplicar', and 'Avançado'. The 'Ferramentas de Dados' group has icons for 'Texto para Colunas', 'Teste de Hipóteses', 'Planilha de Previsão', and 'Previsão'. The 'Estrutura de Tópicos' group includes 'Agrupar', 'Desagrupar', 'Subtotal', and 'Análise'. The 'Análise de Dados' and 'Solver' buttons are also visible. The worksheet area shows a grid from A1 to Y21, with cell H10 currently selected. A small green rectangular selection box is placed over cell I10.

FERRAMENTAS PRÉ-PROCESSAMENTO

Exercício: Abra o arquivo que você limpou antes no Excel e diagnostique o que está ocorrendo. Como você solucionaria?



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
43	3,73E+09	451551	003713-98	297714	/2014/8/3 applicatio	#####	ARQUIVAI	2014	M	Contrato (NULL	arcit	34777.6	PREFEITUF	4,58E+13	G & S TRAI	1,88E+13	AutuaÃ§Ã£o formalizada em decorrÃ¢ncia da r.					
44	3,73E+09	451553	003713-98	297717	/2014/8/3 applicatio	#####	ARQUIVAI	2014	M	Contrato (NULL	arcit	45545.5	PREFEITUF	4,58E+13	WESLEY DI	1,95E+13	AutuaÃ§Ã£o formalizada em decorrÃ¢ncia da r.					
45	1,78E+09	489630	083 e 084-	319275	/2014/4/1 applicatio	#####	ARQUIVAI	2013	M	Contrato (NULL	sberaldo	952865.44	PREFEITUF	5,33E+13	RGO - INC	7,29E+12	Tomada de PreÃ§os nÃº 06/2013 - Contrato 80/2					
46	VigÃ³rcia: 4 meses."																					
47	3,36E+08	512017	025a030-E	336663	/2014/1/3 applicatio	#####	ARQUIVAI	2013	E	Contrato (NULL	robsonma	306000	INSTITUTO	6,18E+13	DIVENA LI	2,99E+12	Edital PE					
48	VigÃ³rcia: 14/06/2013 a 31/12/2013."																					
49	3,37E+08	512018	025a030-E	336666	/2014/1/3 applicatio	#####	ARQUIVAI	2013	E	Contrato (NULL	robsonma	105901.8	INSTITUTO	6,18E+13	FCA FIAT C	1,67E+13	Edital PE					
50	VigÃ³rcia 14/06/2013 a 31/12/2013 -																					
51	PROCESSO PRINCIPAL - eTC 335.989.14-1."																					
52	3,41E+08	512019	025a030-E	336669	/2014/1/3 applicatio	#####	ARQUIVAI	2013	E	Contrato (NULL	robsonma	307200	INSTITUTO	6,18E+13	VOLKSWA	5,91E+13	Edital PE					
53	VENCIMENTO: 14/06/2013 a 31/12/2013 -																					
54	PROCESSO PRINCIPAL: eTC 335.989.14-1."																					
55	3,43E+08	512020	025a030-E	336672	/2014/1/3 applicatio	#####	ARQUIVAI	2013	E	Contrato (NULL	robsonma	153000	INSTITUTO	6,18E+13	EMPORIUI	5,16E+12	Edital PE					
56	VigÃ³rcia: 14/06/2013 a 31/12/2013 -																					
57	PROCESSO PRINCIPAL: eTC 335.989.14-1."																					
58	3,46E+08	512021	025a030-E	336675	/2014/1/3 applicatio	#####	ARQUIVAI	2013	E	Contrato (NULL	robsonma	98620	INSTITUTO	6,18E+13	OK - DISTF	7,89E+12	Edital PE					
59	VigÃ³rcia: 14/06/13 a 31/12/2013 -																					
60	PROCESSO PRINCIPAL: eTC 335.989.14-1."																					
61	1,07E+09	512040	087e088-N	336693	/2014/3/1 applicatio	#####	ARQUIVAI	2014	M	Contrato (NULL	robsonma	10437857	PREFEITUF	4,66E+13	STAFF'S RI	9,06E+09	Edital					
62	4,07E+09	527743	018-E-RM	346993	/2014/9/4 applicatio	#####	ARQUIVAI	2014	E	Contrato (NULL	sberaldo	8104246.3	COMPANH	4,38E+13	ERCON EN	1,72E+13	LICITAÃ‡					
63	2,66E+09	530795	2655.989.1	348910	/2014/6/2 applicatio	#####	ARQUIVAI	2014	E	Contrato (NULL	rmartinsco	4150000	COMPANH	4,38E+13	GUARUJA	8,17E+13	LICITAÃ‡					
64	1,48E+09	549037	04-05-E-DI	359802	/2014/3/1 applicatio	#####	ARQUIVAI	2014	E	Contrato (NULL	erodrigue	937499.55	CENTRO R	4,46E+12	AMBIENT	5,51E+11	Processo					
65	2,23E+09	549062	14-15-M-C	359823	/2013/5/2 applicatio	#####	ARQUIVAI	2013	M	Contrato (NULL	erodrigue	345000	PREFEITUF	4,66E+13	MAESTRO	9,24E+12	EDITAL					
66	4,87E+09	556156	016-E-RM	364379	/2014/10/ applicatio	#####	ARQUIVAI	2014	E	Contrato (NULL	sberaldo	7242648	COMPANH	4,38E+13	F.G.S. BRA	2,29E+12	LICITAÃ‡					
67	1,23E+09	559061	008a010-E	366474	/2014/3/1 applicatio	#####	ARQUIVAI	2014	E	Contrato (REGULARI	robsonma	45936560.	DEPARTAM	4,31E+13	TCL - TECN	4,37E+11	Edital					
68	1,27E+09	559062	008a010-E	366477	/2014/3/1 applicatio	#####	ARQUIVAI	2014	E	Contrato (REGULARI	robsonma	39353006.	DEPARTAM	4,31E+13	CONSTRU	6,11E+13	Edital					
69	6,44E+08	573138	065e066-N	375084	/2014/2/6 applicatio	#####	ARQUIVAI	2012	M	Contrato (IRREGULA	robsonma	21500000	PREFEITUF	4,65E+13	BANCO BF	6,07E+13	Contrato					

FERRAMENTAS PRÉ-PROCESSAMENTO

Exercício: Busque a Lei 14133/2021 na imprensa nacional.

<https://www.in.gov.br/en/web/dou/-/lei-n-14.133-de-1-de-abril-de-2021-311876884>

Clique o botão direito do mouse e selecione “exibir código-fonte”.

CTRL+U

```
<div class="texto-dou"><html>
<head></head>
<body>
<p class="identifica">LEI Nº 14.133, DE 1º DE ABRIL DE 2021</p>
<p class="ementa">Lei de Licitações e Contratos Administrativos.</p>
<p class="dou-paragraph" ><strong class="dou-strong" >O PRESIDENTE DA REPÚBLICA</strong></p>
<p class="dou-paragraph" >Faço saber que o Congresso Nacional decreta e eu sanciono a seguinte Lei:</p>
<p class="dou-paragraph" >TÍTULO I</p>
<p class="dou-paragraph" >DISPOSIÇÕES PRELIMINARES</p>
<p class="dou-paragraph" >CAPÍTULO I</p>
<p class="dou-paragraph" >DO ÂMBITO DE APLICAÇÃO DESTA LEI</p>
<p class="dou-paragraph" >Art. 1º Esta Lei estabelece normas gerais de licitação e contratação para as Administrações Públicas diretas, autárquicas e fundacionais da União, dos Estados, do Distrito Federal e dos Municípios, e abrange as entidades controladas diretamente ou indiretamente pelo Poder Legislativo dos Municípios, quando no desempenho de função administrativa;</p>
<p class="dou-paragraph" >I - os órgãos dos Poderes Legislativo e Judiciário da União, dos Estados e do Distrito Federal e os órgãos do Poder Legislativo dos Municípios, quando no desempenho de função administrativa;</p>
<p class="dou-paragraph" >II - os fundos especiais e as demais entidades controladas diretamente ou indiretamente pela Administração Pública.</p>
<p class="dou-paragraph" >§ 1º Não são abrangidas por esta Lei as empresas públicas, as sociedades de economia mista e as suas subsidiárias, regidas pela Lei nº 13.303, de 30 de junho de 2016, ressalvado o disposto no art. 178 da Constituição Federal.</p>
<p class="dou-paragraph" >§ 2º As contratações realizadas no âmbito das repartições públicas sediadas no exterior obedecerão às peculiaridades locais e aos princípios básicos estabelecidos nesta Lei, na forma de regulamentação específica.</p>
<p class="dou-paragraph" >§ 3º Nas licitações e contratações que envolvam recursos provenientes de empréstimo ou doação oriundos de agência oficial de cooperação estrangeira ou de organismo financeiro de que o Brasil seja parte, as condições decorrentes de acordos internacionais aprovados pelo Congresso Nacional e ratificados pelo Presidente da República;</p>
<p class="dou-paragraph" >I - condições peculiares à seleção e à contratação constantes de normas e procedimentos das agências ou dos organismos, desde que:</p>
<p class="dou-paragraph" >a) sejam exigidas para a obtenção do empréstimo ou doação;</p>
<p class="dou-paragraph" >b) não conflitem com os princípios constitucionais em vigor;</p>
<p class="dou-paragraph" >c) sejam indicadas no respectivo contrato de empréstimo ou doação e tenham sido objeto de parecer favorável do órgão jurídico do contratante do financiamento previamente à celebração do referido contrato;</p>
<p class="dou-paragraph" >d) (VETADO).</p>
<p class="dou-paragraph" >§ 4º A documentação encaminhada ao Senado Federal para autorização do empréstimo de que trata o § 3º deste artigo deverá fazer referência às condições contratuais que incidam na hipótese do referido contrato.</p>
<p class="dou-paragraph" >§ 5º As contratações relativas à gestão, direta e indireta, das reservas internacionais do País, inclusive as de serviços conexos ou acessórios a essa atividade, serão disciplinadas em ato normativo próprio.</p>
<p class="dou-paragraph" >Art. 2º<strong class="dou-strong" >Esta Lei aplica-se a:</strong></p>
<p class="dou-paragraph" >I - alienação e concessão de direito real de uso de bens;</p>
<p class="dou-paragraph" >II - compra, inclusive por encomenda;</p>
<p class="dou-paragraph" >III - locação;</p>
<p class="dou-paragraph" >IV - concessão e permissão de uso de bens públicos;</p>
<p class="dou-paragraph" >V - prestação de serviços, inclusive os técnico-profissionais especializados;</p>
<p class="dou-paragraph" >VI - obras e serviços de arquitetura e engenharia;</p>
<p class="dou-paragraph" >VII - contratações de tecnologia da informação e de comunicação.</p>
<p class="dou-paragraph" >Art. 3º Não se subordinam ao regime desta Lei:</p>
<p class="dou-paragraph" >I - contratos que tenham por objeto operação de crédito, interno ou externo, e gestão de dívida pública, incluídas as contratações de agente financeiro e a concessão de garantia relacionadas a esses contratos;</p>
<p class="dou-paragraph" >II - contratações sujeitas a normas previstas em legislação própria.</p>
<p class="dou-paragraph" >Art. 4º Aplicam-se as licitações e contratos disciplinados por esta Lei as disposições constantes dos arts. 42 a 49 da Lei Complementar nº 123, de 14 de dezembro de 2006.</p>
<p class="dou-paragraph" >§ 1º As disposições a que se refere o caput deste artigo não são aplicadas:</p>
<p class="dou-paragraph" >I - no caso de licitação para aquisição de bens ou contratação de serviços em geral, ao item cujo valor estimado for superior à receita bruta máxima admitida para fins de enquadramento como empresa de pequeno porte;</p>
<p class="dou-paragraph" >II - no caso de contratação de obras e serviços de engenharia, às licitações cujo valor estimado for superior à receita bruta máxima admitida para fins de enquadramento como empresa de pequeno porte.</p>
<p class="dou-paragraph" >§ 2º A obtenção de benefícios a que se refere o caput deste artigo fica limitada às microempresas e às empresas de pequeno porte que, no ano-calendário de realização da licitação, ainda não tenham celebrado contratos com prazo de vigência superior a 1 (um) ano, será considerado o valor anual do contrato na aplicação dos limites previstos nos §§ 1º e 2º deste artigo.</p>
<p class="dou-paragraph" >CAPÍTULO II</p>
```

FERRAMENTAS PRÉ-PROCESSAMENTO

Escolha “salvar como”, opção html.

Abra o arquivo no Notepad++.

Qual é a codificação?

Qual é o identificador de final de linha?

FERRAMENTAS PRÉ-PROCESSAMENTO

Volte na lei 14.133/2021 do site da imprensa nacional.

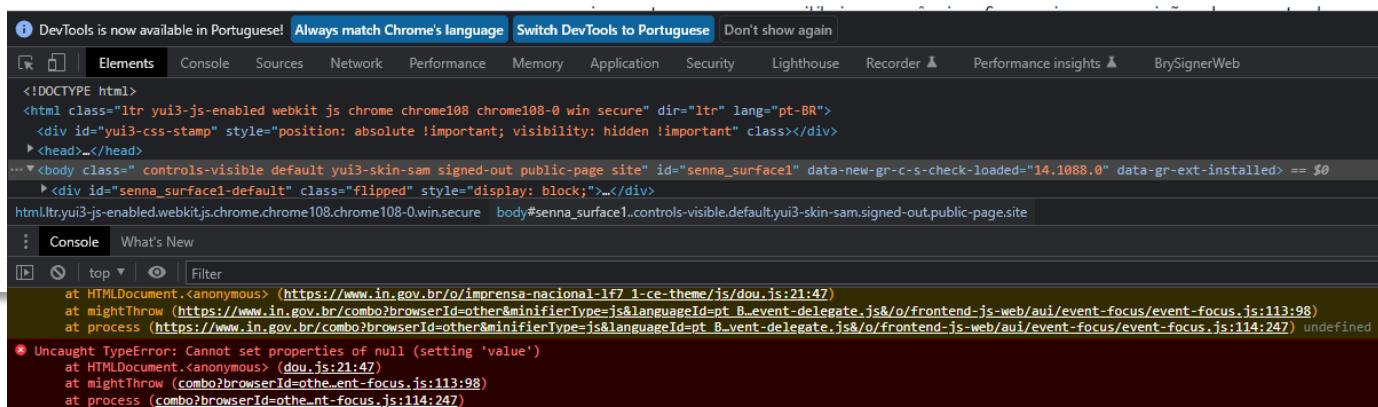
Acesse as ferramentas do desenvolvedor (CTRL+SHIFT+I)

nos incisos I, II, III, IV e VII do **caput** do art. 4º desta Lei;

XXVI - projeto executivo: conjunto de elementos necessários e suficientes à execução completa da obra, com o detalhamento das soluções previstas no projeto básico, a identificação de serviços, de materiais e de equipamentos a serem incorporados à obra, bem como suas especificações técnicas, de acordo com as normas técnicas pertinentes;

XXVII - matriz de riscos: cláusula contratual definidora de riscos e de responsabilidades entre as partes e caracterizadora do equilíbrio econômico-financeiro inicial do contrato, em termos de ônus financeiro decorrente de eventos supervenientes à contratação, contendo, no mínimo, as seguintes informações:

a) listagem de possíveis eventos supervenientes à assinatura do contrato que possam causar



The screenshot shows the developer tools of a web browser. The title bar indicates "DevTools is now available in Portuguese! Always match Chrome's language Switch DevTools to Portuguese Don't show again". The main area displays the DOM tree, which includes the DOCTYPE, HTML, head, and body sections. The body section has an ID of "senna_surface1" and a class of "controls-visible default yui3-skin-sam signed-out public-page site". Below the DOM tree, the "Console" tab is active, showing several error messages in yellow and red. One error message at the bottom is highlighted in red: "Uncaught TypeError: Cannot set properties of null (setting 'value') at HTMLDocument.<anonymous> (https://www.in.gov.br/o/imprensa-nacional-1f71-ce-theme/js/dou.js:21:47)". Other errors mention "mightThrow" and "process" functions.

FERRAMENTAS PRÉ-PROCESSAMENTO

Acesse o console:

```
var txConteudo = document.body.textContent
// se for no Edge var txConteudo = document.body.innerText
function copiarClipboard(texto) {
    var dummy = document.createElement("textarea");
    // a linha abaixo evita um erro que eu não me lembro
    document.body.appendChild(dummy);
    dummy.value = texto;
    dummy.select();
    document.execCommand("copy");
    document.body.removeChild(dummy);
}
copiarClipboard(txConteudo)
```

FERRAMENTAS PRÉ-PROCESSAMENTO

Agora copie no word.

Como isso é diferente de selecionar o texto e efetuar CTRL+C e CTRL+V?

Qual a vantagem?

Usar o Document Object Model da página HTML

Há um plugin do chrome:

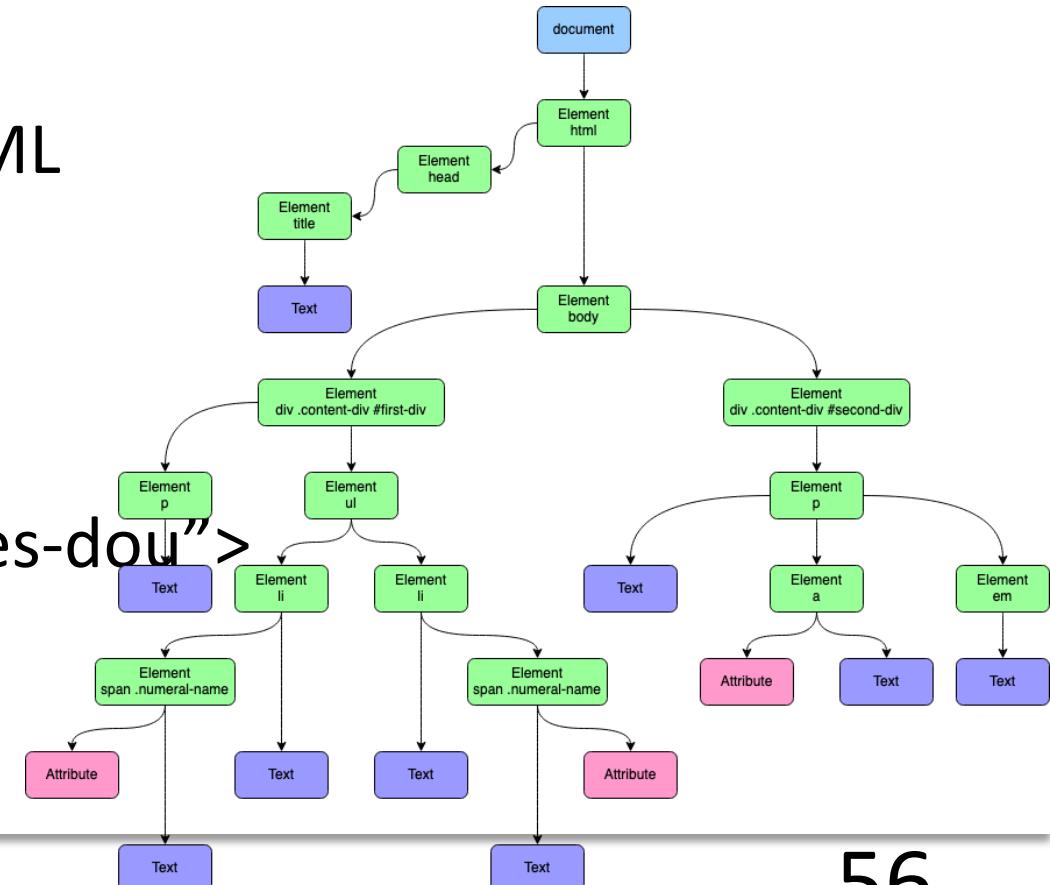
HTML Tree Generator

E se eu desejar o conteúdo de apenas um div?

<div class="“texto-dou”> ou <div class="“detalhes-dou”>

E se eu desejar remover o que estiver taxado?

<s>tachado</s>



FERRAMENTAS PRÉ-PROCESSAMENTO

Veja alguma referência, tal como:

https://fundamentals.generalassemb.ly/11_unit/dom-cheatsheet.html

Method	Description
getElementById()	Selects an individual element within a document using a specific id
querySelector()	Uses CSS selector to select the first matching element within a document
getElementsByClassName()	Allows you to select all elements with a given class attribute
getElementsByTagName()	Locates all elements that match a given tag name
querySelectorAll()	Uses CSS selector to select one or more elements

FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX

“Expressões regulares” são uma sublinguagem empregada para escolher intervalos de códigos de um arquivo texto.

No notepad++ podem ser utilizadas também para identificar, contar, extrair ou substituir.

Os comandos REGEX são identificados com uma barra invertida “\”.

Por exemplo \n localiza o LF, \r localiza o CR, \t localiza o TAB.

Se escolhermos, por exemplo, substituir todos os duplos fins de linha por um único fim de linha (textos importados do site do planalto vem com duplo fim de linha), substituiríamos “\n\r\n\r” por “\n\r”.

FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX

Expressões literais [a-z]

“Grupos” ()

Look Around – positive look behind (?<=a) negative look behind (?<!a)

Look Around – positive look ahead (?=a) negative look ahead (?!a)

Numerosidade – * 0 ou mais, + 1 ou mais, ? Operador “ganancioso”

\s – espaço em branco

\d – dígito

\w – imprimível “word”

\b – início e fim de uma sequência de imprimíveis

^ - no início da linha (logo após de \n\r)

\$ - ao final da linha (logo antes de \n\r)

FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX - ferramentas

<https://www.regextester.com/>

<https://www.regular-expressions.info/>

<https://www.rexegg.com/>

<https://www.regexpal.com/>

<https://www.regexmagic.com/>

E, é claro, o Stackoverflow!!

<https://stackoverflow.com/>

FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX

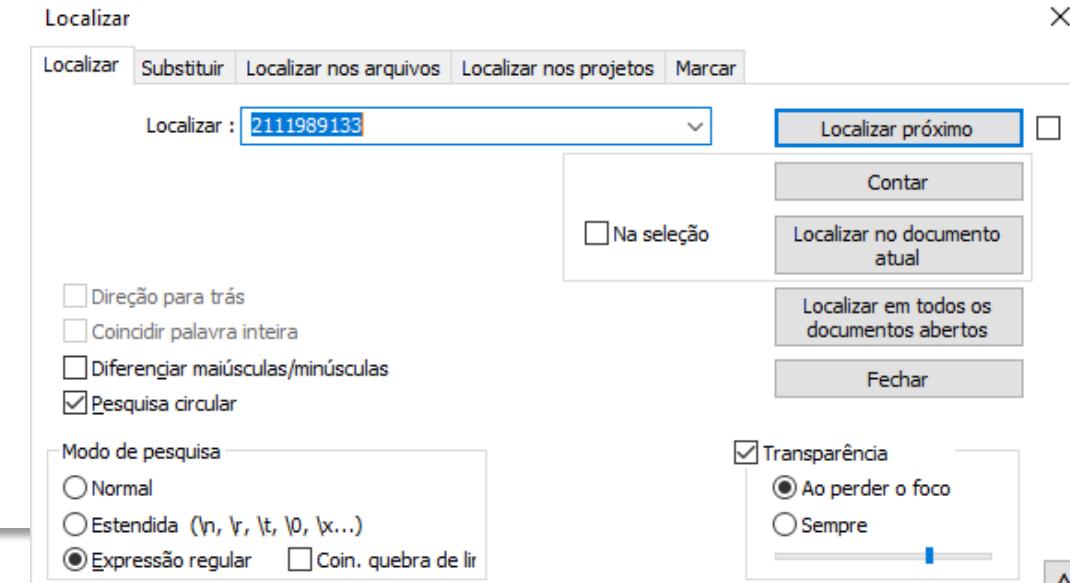
Ferramentas do Notepad++:

Para exibir os elementos não imprimíveis, use Visualizar>Exibir todos os caracteres

Para localizar uma expressão regular, use Localizar>Localizar (CTRL+F)

Em “modo de pesquisa”,

escolha “expressão regular”.



FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX

Exercício: abra novamente o arquivo que você limpou “contratos etcesp v1 poluído.csv”

Tente elaborar um regex para contar a quantidade de referências a um número de processo.

São até nove dígitos para o número (ao menos um, pois não há processo “0”), seguidos por “989”, seguidos por exatamente três dígitos.

1. Como você faria para contar quantas ocorrências são?
2. Como você faria para copiar todas as ocorrências para um arquivo Excel?
3. Como você faria para excluir todas as ocorrências?
4. Como você faria para mudar todas as ocorrências para o formato xxx.xxx/989/xx-x?

FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX

Tente elaborar um regex para contar a quantidade de referências a um número de processo.

\d{1,9}989\d{3}

1. Como você faria para contar quantas ocorrências são?

CTRL+F > “Contar”

2. Como você faria para copiar todas as ocorrências para um arquivo Excel?

CTRL+F > “Marcar” > “Localizar todos” > “Copiar texto marcado”

Ou Localizar > “Marcadores” > “Copiar Linhas com Marcadores”, para pegar a linha.

Após, abrir o excel e escolher o colar especial > texto UNICODE.

3. Como você faria para excluir todas as ocorrências?

CTRL+F > “Substituir” (Ou CTRL+H) > Apagar o que constar em “Substituir por”

4. Como você faria para mudar todas as ocorrências para o formato xxxxxx/989/xx-x?

CTRL+H > Em localizar: “(\d{1,9})(989)(\d{2})(\d{1})” Em Substituir por “\1\2\3\-\4”

FERRAMENTAS PRÉ-PROCESSAMENTO

REGEX

Em nosso arquivo “contratos etcesp v1 poluído.csv” há dois problemas:
o campo “resumo” contém quebras de linha.
aspas (o caracter ”) é duplicado.

Estas duas ocorrências são frequentes quando há trânsito pelo Excel.

Como substituiríamos essas quebras de linha, por exemplo, por um espaço
(concatenar todo o texto em uma única linha?)

O que faz a seguinte expressão?

```
\r\n(?!d{1,9}989)d{3}
```

O que ocorre se substituíssemos todas essas ocorrências por um espaço em branco?

FERRAMENTAS PRÉ-PROCESSAMENTO

^(CAPÍTULO)(.+)\r\n(.+)

^(^ \t)+.*

\1

^(Seção
)(M{0,4}(CM|CD|D?C{0,3})(XC|XL|L?X{0,3})(IX|IV|
V?I{0,3}))\r\n\n
\1\2

^§\s
§

\(Redação.+?(?=\\))\)

\(Incluído.+?(?=\\))\)

\(Revogado .+?(?=\\))\)

identificacao das notas de rodape

\r\n([\d]+\s)([A-Z])
\r\n###rodape###\1\2

identificacao dos dispositivos numericos

\r\n(\d+.+)
\r\n###dispositivo_num###\1

identificacao dos dispositivos letra

\r\n([a-z]+\\))
\r\n###dispositivo_letra###\1

trocar tabs por espacos

\t
" "

identificacao de incisos

^([LXVI]+\\s)
###dispositivo_inciso###\1

identificacao dos capitulos, seções e subseções

^CAPÍTULO
###dispositivo_capitulo###CAPÍTULO

^SEÇÃO
###dispositivo_secao###SEÇÃO

^Subseção
###dispositivo_subseção###Subseção

retirar as quebras de linha excedentes

\r\n(^#)
\1

FERRAMENTAS PRÉ-PROCESSAMENTO

Entre colchetes	\{.*\}
HTML	<(\V \\\)?.+?>
Números	\d
Email	[^\s]+@[^\s]+
URL	(http https)://[^\s]+
Aspas finais	(^ \W)[“”]
Aspas iniciais	[“”](\W \\$)
Aspas soltas	[“”””]
Dois pontos soltos	(?<!\.)\.\.(?!\.)
Pontuação	([,":])\{2},
Hífens	\s-(?=^[^\W\d_])
Aspas duplas em duplicidade	(\"\\")
Aspas simples em duplicidade	(\'\\')
Espaços	\s+
Introduzir espaços em branco entre palavra e pontuação	(\w+)([!"#%\(')*+,./;:<=>?@\[\]\^_\`{ }~])([\s!"#%\(')*+,./;:<=>?@\[\]\^_\`{ }~])
Introduzir espaços em branco entre palavra e pontuação	([\s!"#%\(')*+,./;:<=>?@\[\]\^_\`{ }~])([!"#%\(')*+,./;:<=>?@\[\]\^_\`{ }~])(\w+)
Introduzir espaços em branco entre palavra e pontuação	(\w+)([!"#%\(')*+,./;:<=>?@\[\]\^_\`{ }~])\\$

FERRAMENTAS PRÉ-PROCESSAMENTO

POSTGRES

Acessamos o Postgres a partir do python com o psycopg3 (no conda-forge só tem o psycopg2)

```
import psycopg
```

```
with psycopg.connect("dbname=postgres user=postgres password=postgres") as conn:  
    with conn.cursor() as cur:  
        cur.execute("""CREATE TABLE paragrafos (id serial PRIMARY KEY, ord integer, texto text,  
                    vetor double precision ARRAY[%s] """", (dim_embedding))  
        cur.execute("INSERT INTO paragrafos (ord, text, vetor) VALUES (%s, %s, %s)", \  
                   (100, texto[100], embedding[100]))  
        cur.execute("SELECT * FROM paragrafos")  
        cur.fetchone()  
        for record in cur:  
            print(record)  
        conn.commit()
```

FERRAMENTAS PRÉ-PROCESSAMENTO

POSTGRES

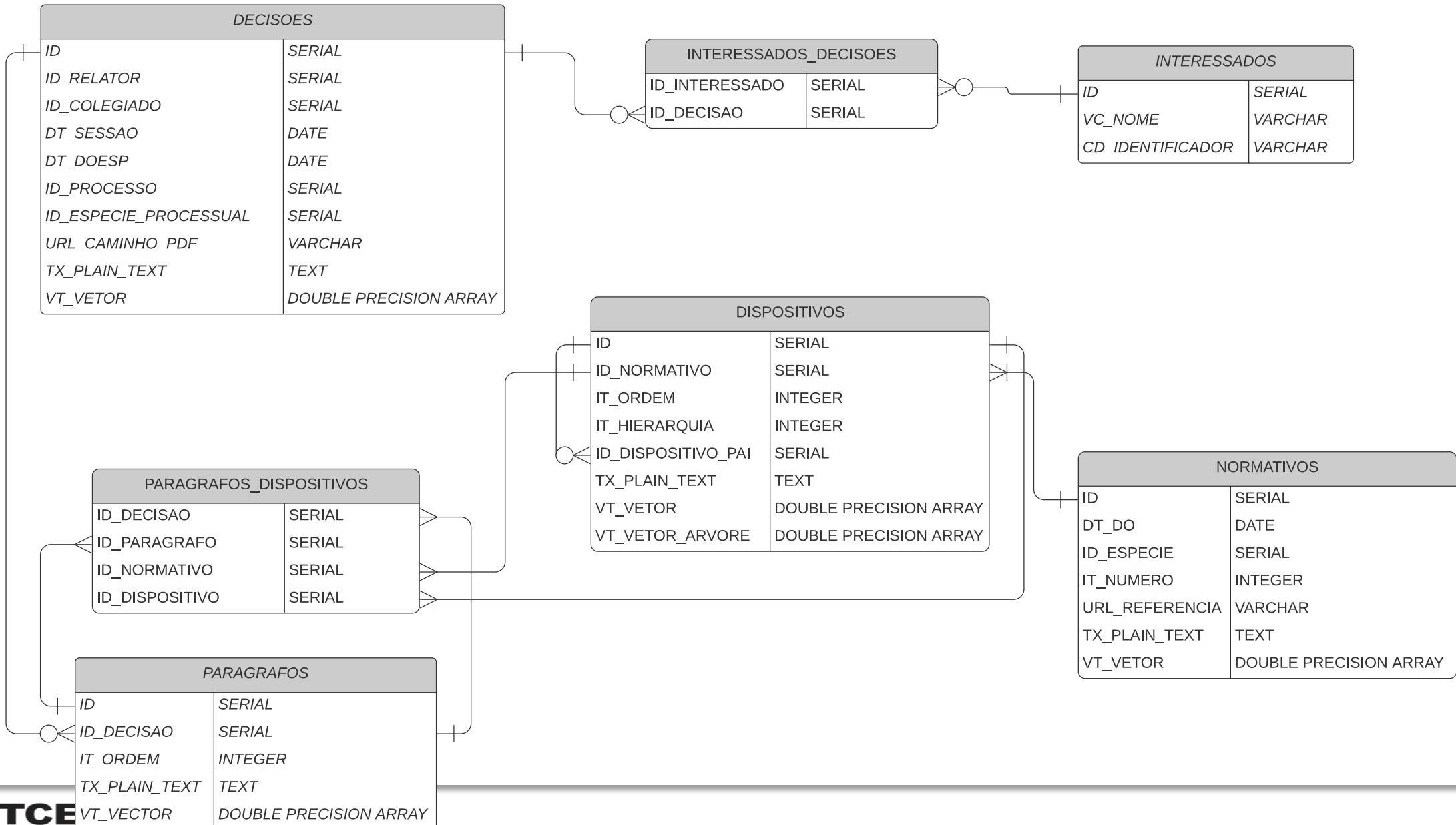
Acessamos o Postgres a partir do python com o psycopg3 (no conda-forge só tem o psycopg2)

```
import psycopg
```

```
with psycopg.connect("dbname=postgres user=postgres password=postgres") as conn:  
    with conn.cursor() as cur:  
        cur.execute("""CREATE TABLE paragrafos (id serial PRIMARY KEY, ord integer, texto text,  
                    vetor double precision ARRAY[%s] """", (dim_embedding))  
        cur.execute("INSERT INTO paragrafos (ord, text, vetor) VALUES (%s, %s, %s)", \  
                   (100, texto[100], embedding[100]))  
        cur.execute("SELECT * FROM paragrafos")  
        cur.fetchone()  
        for record in cur:  
            print(record)  
        conn.commit()
```

DIAGRAMA ER - NLP decisões

Alexandre Sarquis | December 10, 2022



FERRAMENTAS PRÉ-PROCESSAMENTO

O que é pré-processar?

Pode significar diferentes coisas para diferentes pessoas (diferentes pipelines)

Deixar tudo minúsculo?

Excluir “stop-words”?

Separar palavras (tokens)?

Juntar palavras (named-entities)?

Lematizar?

Excluir espaços?

Incluir espaços?

É produzir um formato específico de arquivo?

É popular uma tabela de banco de dados relacional, incluindo metadados?

As diversas ferramentas apresentam diversas velocidades, diversas proficiências e diversas clarezas para diversas funções.

A ferramenta mais geral é o python.

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
- 3. REPRESENTAÇÃO DE TEXTO PARA NLP**
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

REPRESENTAÇÃO DE TEXTO

Temos a linguagem – será a geometria analítica.

Temos o texto pré-processado.

Como partimos de um texto e chegamos em um tensor?

REPRESENTAÇÃO DE TEXTO

- 1. TOKEN
- 2. BAG OF WORDS
- 3. WORDCLOUD
- 4. WORD
EMBEDDING
- 5. CBOW
- 6. SKIPGRAM
- 7. WORD2VEC
- 8. GLOVE
- 9. INFORMAÇÃO
- 10. TF-IDF
- 11. OKAPI BM25

REPRESENTAÇÃO DE TEXTO

“Conheço dos embargos de declaração opostos para, no mérito, rejeitá-los.”

“Em preliminar, conheço, mas, no mérito, rejeito.”

“Nego provimento.”

REPRESENTAÇÃO DE TEXTO

“Conheço dos embargos de declaração
opostos para, no mérito, rejeitá-los.”

Stop words – dos, de, para, no, los.

Named Entity – embargos_de_declaracão

Lemmas – conhecer, opor, mérito, rejeitar

REPRESENTAÇÃO DE TEXTO

“Conheço dos embargos de declaração opostos para, no mérito, rejeitá-los.”

Verbos – conheço, opostos, rejeitá-los.

Substantivos – embargos de declaração, mérito, eu (oculto).

Preposições – dos, de, no, para.

REPRESENTAÇÃO DE TEXTO

LEMMA	OCORRÊNCIA	TF-IDF
CONHECER	80%	CONHECER
EMBARGOS_DE_DECLARAÇÃO	10%	EMBARGOS_DE_DECLARAÇÃO
OPOR	30%	OPOR
MÉRITO	90%	MÉRITO
REJEITAR	14%	REJEITAR

REPRESENTAÇÃO DE TEXTO

```
import spacy
```

```
text = """Conheço dos embargos de declaração opostos para, no  
mérito, rejeitá-los."""
```

```
nlp = spacy.load("pt_core_news_lg")
```

```
doc = nlp(text)
```

```
for token in doc:  
    print(token.text, token.pos_, token.dep_, token.lemma_, \  
          token.tag_, token.dep_, token.shape_, \  
          token.is_alpha, token.is_stop, token.has_vector, \  
          token.vector_norm, token.is_oov)
```

REPRESENTAÇÃO DE TEXTO

text	pos	dep	lemma	tag	dep	shape	is_alpha	is_stop	has_vector	vector_norm	is_oov
Conheço	VERB	ROOT	Conheço	VERB	ROOT	Xxxxx	True	False	True	20	False
dos	ADP	case	De o	ADP	case	xxx	True	True	True	61	False
embargos	NOUN	obj	embargo	NOUN	obj	xxxx	True	False	True	31	False
de	ADP	case	de	ADP	case	xx	True	True	True	61	False
declaração	NOUN	nmod	declaração	NOUN	nmod	xxxx	True	False	True	28	False
opostos	ADJ	amod	oposto	ADJ	amod	xxxx	True	False	True	27	False
para	SCONJ	mark	para	SCONJ	mark	xxxx	True	True	True	35	False
,	PUNCT	punct	,	PUNCT	punct	,	False	False	True	34	False
no	ADP	case	em o	ADP	case	xx	True	True	True	81	False
mérito	NOUN	obl	mérito	NOUN	obl	xxxx	True	False	True	28	False
,	PUNCT	punct	,	PUNCT	punct	,	False	False	True	34	False
rejeitá-los	VERB	advcl	rejeitá-los	VERB	advcl	xxxx-xxx	False	False	True	20	False
.	PUNCT	punct	.	PUNCT	punct	.	False	False	True	44	False

REPRESENTAÇÃO DE TEXTO

spacy.io/api/token

spaCy Out now: spaCy v3.4

OVERVIEW

- Library Architecture
- Model Architectures
- Data Formats
- Command Line
- Functions

CONTAINERS

- Doc
- DocBin
- Example
- Language
- Lexeme
- Span
- SpanGroup
- Token

PIPELINE

- AttributeRuler
- CoreferenceResolver
- DependencyParser
- EditTreeLemmatizer
- EntityLinker
- EntityRecognizer
- EntityRuler
- Lemmatizer
- Morphologizer
- SentenceRecognizer
- Sentencizer
- SpanCategorizer
- SpanResolver
- SpanRuler
- Tagger
- TextCategorizer
- Tok2Vec

	TYPE:	Span
text	TYPE: str	Verbatim text content.
text_with_ws	TYPE: str	Text content, with trailing space character if present.
whitespace_	TYPE: str	Trailing space character if present.
orth	TYPE: int	ID of the verbatim text content.
orth_	TYPE: str	Verbatim text content (identical to <code>Token.text</code>). Exists mostly for consistency with the other attributes.
vocab	TYPE: vocab	The vocab object of the parent <code>Doc</code> .
tensor	TYPE: numpy.ndarray	The token's slice of the parent <code>Doc</code> 's tensor.
head	TYPE: Token	The syntactic parent, or "governor", of this token.
left_edge	TYPE: Token	The leftmost token of this token's syntactic descendants.
right_edge	TYPE: Token	The rightmost token of this token's syntactic descendants.
i	TYPE: int	The index of the token within the parent document.
ent_type	TYPE: int	Named entity type.
ent_type_	TYPE: str	Named entity type.

REPRESENTAÇÃO DE TEXTO

“Conheço dos embargos de declaração opostos para, no mérito, rejeitá-los.”

<http://conteudo.icmc.usp.br/pessoas/taspardo/LematizadorV2a.rar>

Thiago A. S. Pardo – USP São Carlos

java –jar lematizador argumento1 argumento2

REPRESENTAÇÃO DE TEXTO

```
D:\Users\Alexandre\Documents\github\NLP\LEMATIZADOR>java -jar lematizador.jar input.txt nf
Wait...
0% completed...
4% completed...
8% completed...
12% completed...
16% completed...
20% completed...
25% completed...
29% completed...
33% completed...
37% completed...
41% completed...
45% completed...
50% completed...
54% completed...
58% completed...
62% completed...
66% completed...
70% completed...
75% completed...
79% completed...
83% completed...
87% completed...
91% completed...
95% completed...
paragraph...
Finished!
```

```
D:\Users\Alexandre\Documents\github\NLP\LEMATIZADOR>
```

REPRESENTAÇÃO DE TEXTO

HABEAS CORPUS.. DIREITO PENAL E DIREITO PROCESSUAL PENAL.. ENTIDADE FECHADA DE PREVIDÃ‰NCIA PRIVADA.. EQUIPARAÃ‰SÃ‰fO A INSTITUIÃ‰SÃ‰fO FINANCEIRA.. GESTÃ‰fO FRAUDULENTA..
1.. InstituiÃ§Ã£o financeiro, para o fim do Lei nÃº 7.492 / 86, Ã© todo e qualquer pessoa jurÃ©dica de direito pÃºblico ou privado, que, como atividade principal
2.. o que caracterizar, para o fim do Lei nÃº 7.492 / 86, o instituiÃ§Ã£o financeiro, de natureza pÃºblica ou privado, Ã©, essencialmente, que o ele atividade,,
3.. o entidade fechado de previdÃ‰ncia privado, que capta e administrar recurso destinado ao pagamento de benefÃºcio de ele associado, equiparar (-se[enc]) o instituiÃ§Ã£o
4.. o fato de estatuir o Lei nÃº 4.565 / 64, no letra de ele artigo 25, com o redaÃ§Ã£o que ele ser/ir atribuÃ‰do pelo Lei nÃº 5.710, de 7 de outubro de 1971, que
5.. quando ele negar que o entidade fechado de previdÃ‰ncia privado nÃ£o participa do natureza do instituiÃ§Ã£o es de seguro, o disposiÃ§Ã£o que inserta no inciso I do par
6.. desse modo, por forÃ§Ão de natureza ou pelo equiparaÃ§Ã£o levar o cabo pelo Lei 7.492 / 86 (artigo 1Ãº, parÃ¡grafo 1º, inciso I, parte final), nÃ£o hÃ©
7.. o Emenda Constitucional nÃº 40 / 2003 - que reduzir o disposiÃ§Ã£o referente ao Sistema Financeiro Nacional ao que ser/erar o caput do artigo 192, com ligeiro modificac
8.. Ã‰ que o decisÃ£o polÃ¢tica de envio do entidade fechado de previdÃ‰ncia complementar do capÃºtulo prÃ³prio do Sistema Financeiro Nacional para o capÃºtulo do Se
9.. Ã‰ do competÃ‰ncia do JustiÃ§Ãa federal o julgamento do crime contra o Sistema Financeiro Nacional, tipificar no Lei nÃº 7.492 / 86 (artigo 26).. **CRLE**
10.. Em haver crime conexo, apenado diversamente, o competÃ‰ncia para processar e julgar o aÃ§Ão penal Ã© definir pelo lugar do crime cujo pena Ã© mais gravoso, prev
11.. o competÃ‰ncia para processar e julgar o crime conexo tipificar no artigo 4Ãº, 5Ãº, caput, e parÃ¡grafo 1º, 6Ãº, 7Ãº, inciso IV, 9Ãº, 10 e 17, do Lei
12.. contribuir o paciente efetivamente para o prÃ¶tica do evento delituoso, nÃ£o hÃ¡; falar em inexistÃ‰ncia de prova de ele participaÃ§Ãão, tanto quanto nÃ£o hÃ¡;
13.. Ordem denegar.. **CRLE**
(HC 26.288 / SP, Rel. Ministro HAMILTON CARVALHIDO, SEXTA TURMA, julgar em 03 / 02 / 2005, DJ 11 / 04 / 2005, p. 385) **CRLE**
CRLE

REPRESENTAÇÃO DE TEXTO

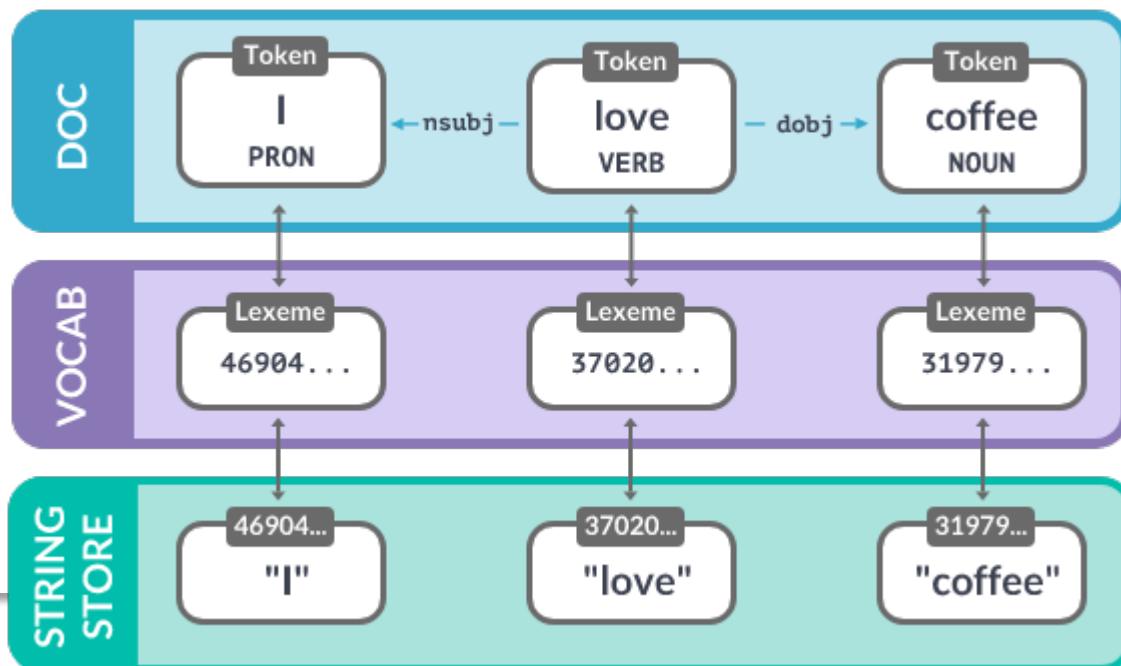
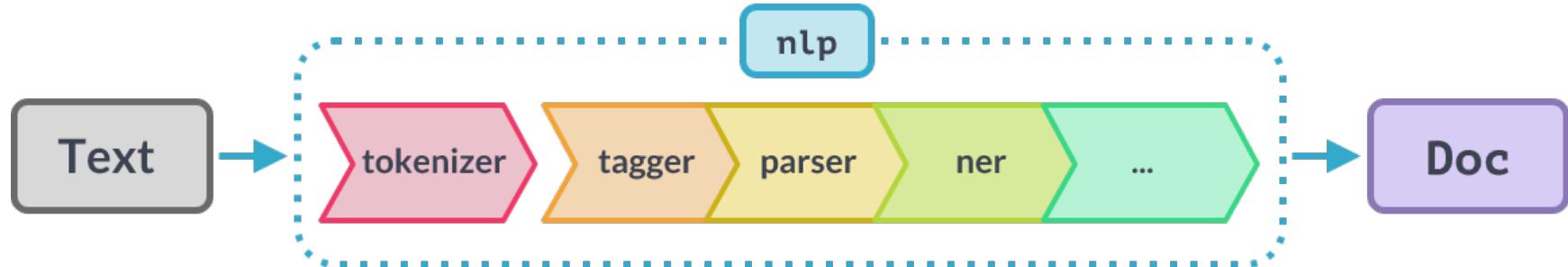
```
D:\Users\Alexandre\Documents\github\NLP\LEMATIZADOR>java -jar lematizador.jar input.txt nf
Wait...
0% completed...
6% completed...
13% completed...
20% completed...
26% completed...
33% completed...
40% completed...
46% completed...
53% completed...
60% completed...
66% completed...
73% completed...
80% completed...
86% completed...
93% completed...
paragraph...
Finished!
```

```
D:\Users\Alexandre\Documents\github\NLP\LEMATIZADOR>
```

conhecer do embargo de declaração oposto para , no mérito , rejeitá (-los[enc]) .

REPRESENTAÇÃO DE TEXTO

Spacy	NILC
Conheço	conhecer
dos	do
embargos	embargo
de	de
declaração	declaração
opostos	oposto
para	para
,	,
no	no
mérito	mérito
,	,
rejeitá-los	rejeitá (-los[enc])
.	.



REPRESENTAÇÃO DE TEXTO

```
$pip install  
https://huggingface.co/explosion/pt\_udv25\_portuguesebosque\_trf/resolve/main/pt\_udv25\_portuguesebosque\_trf-any-py3-none-any.whl
```

```
nlp_trans = spacy.load("pt_udv25_portuguesebosque_trf")  
doc_trans = nlp_trans(text)  
for token in doc_trans:  
    print(token.text, token.pos_, token.dep_, token.lemma_, \  
          token.tag_, token.dep_, token.shape_, \  
          token.is_alpha, token.is_stop, token.has_vector, \  
          token.vector_norm, token.is_oov)
```

REPRESENTAÇÃO DE TEXTO

text	pos	dep	lemma	tag	dep	shape	is_alpha	is_stop	has_vector	vector_norm	is_oov
Conheço	VERB	ROOT	Conheço	VERB	ROOT	Xxxxx	True	False	False	0.0	True
dos	ADP	case	de o	ADP_DET	case	xxx	True	True	False	0.0	True
embargos	NOUN	obj	embargo	NOUN	obj	xxxx	True	False	False	0.0	True
de	ADP	case	de	ADP	case	xx	True	True	False	0.0	True
declaração	NOUN	nmod	declaração	NOUN	nmod	xxxx	True	False	False	0.0	True
opostos	VERB	acl	opostar	VERB	acl	xxxx	True	False	False	0.0	True
para	SCONJ	mark	para	SCONJ	mark	xxxx	True	True	False	0.0	True
,	PUNCT	punct	,	PUNCT	punct	,	False	False	False	0.0	True
no	ADP	case	em o	ADP_DET	case	xx	True	True	False	0.0	True
mérito	NOUN	obl	mérito	NOUN	obl	xxxx	True	False	False	0.0	True
,	PUNCT	punct	,	PUNCT	punct	,	False	False	False	0.0	True
rejeitá-los	VERB	advcl	rejeitá-los	VERB_PRON	advcl	xxxx-xxx	False	False	False	0.0	True
.	PUNCT	punct	.	PUNCT	punct	.	False	False	False	0.0	True

REPRESENTAÇÃO DE TEXTO

O que é “embedding”? “One-hot” encoding

	x_0	x_1	x_2	x_3	x_4	x_5	x_6
Rei	1	0	0	0	0	0	0
Rainha	0	1	0	0	0	0	0
Homem	0	0	1	0	0	0	0
Mulher	0	0	0	1	0	0	0
Cachorro	0	0	0	0	1	0	0
Gato	0	0	0	0	0	1	0
Cachorrão	0	0	0	0	0	0	1

REPRESENTAÇÃO DE TEXTO

Há diversos problemas com a estratégia de “one-hot” encoding

- Ineficiência de armazenamento – matrizes esparsas (mais de 90% de entradas nulas)
- Multiplicidade de pesos
- Insegurança quanto às interpretações de negócio
 - Os tokens “rápido”, “veloz”, “acelerado”, “ligeiro” e “súbito” todos recebem codificações totalmente diversas, e indistinguíveis, a priori, dos tokens “lento”, “vagaroso” e “arrastado”, por exemplo.
- Irrelevância do módulo. O vetor de cada token é unitário sempre. Da mesma forma, não haveria sentido em valores negativos, desperdiçando sentido.
- Se, no exemplo, tivéssemos uma dimensão em que 1 é humano, outra dimensão em que 1 é dominante e outra dimensão em que 1 é feminino, talvez possamos compactar as 6 dimensões originais em 3.

REPRESENTAÇÃO DE TEXTO

Embedding

	x_Humano	x_Dominante	x_Feminino
Rei	1	1	-1
Rainha	1	1	1
Homem	1	0	-1
Mulher	1	0	1
Cachorro	-1	0	-1
Gato	-1	-0,5	0,5
Cachorrão	-1	0,5	-1

REPRESENTAÇÃO DE TEXTO

- A grande questão é como fazer o embedding...
- O usual é realizar o embedding de forma orgânica, ou seja, usando dados não etiquetados (unsupervised learning).
- Técnicas “word2vec” existentes:
 - Skipgram x CBOW
- Confiam na cadeia de Markov
- NÃO identificam sinônimos!!! Identificam – no máximo – palavras que usualmente se apresentam nos mesmos contextos – contexto significando as mesmas palavras adjacentes.
- A dimensão “m” origem é o tamanho do vocabulário, por exemplo 1.000.000.
- A dimensão “n” alvo é uma opção de projeto, tenho visto $n = 100$.
- Já usei dimensão $n = 1000$ sem ganhos muito evidentes.
- As dimensões não terão sentido semântico claro ou a priori.

REPRESENTAÇÃO DE TEXTO

- “Você é muito xxxx.”
- Xxxx costuma ser bem diversa! Mas o word embedding vai levar todas essas palavras diversas a regiões similares do espaço, sejam elas favoráveis, sejam elas desfavoráveis.
- “Bonita” ficará próximo de “Feia”.
- “Skip”-gram: A partir da palavra que se deseja inferir o vetor adequado, verificam-se as vizinhas, calculando a perda, que é minimizada.
- CBOW: A partir das vizinhas àquela que se deseja inferir o vetor adequado.

REPRESENTAÇÃO DE TEXTO

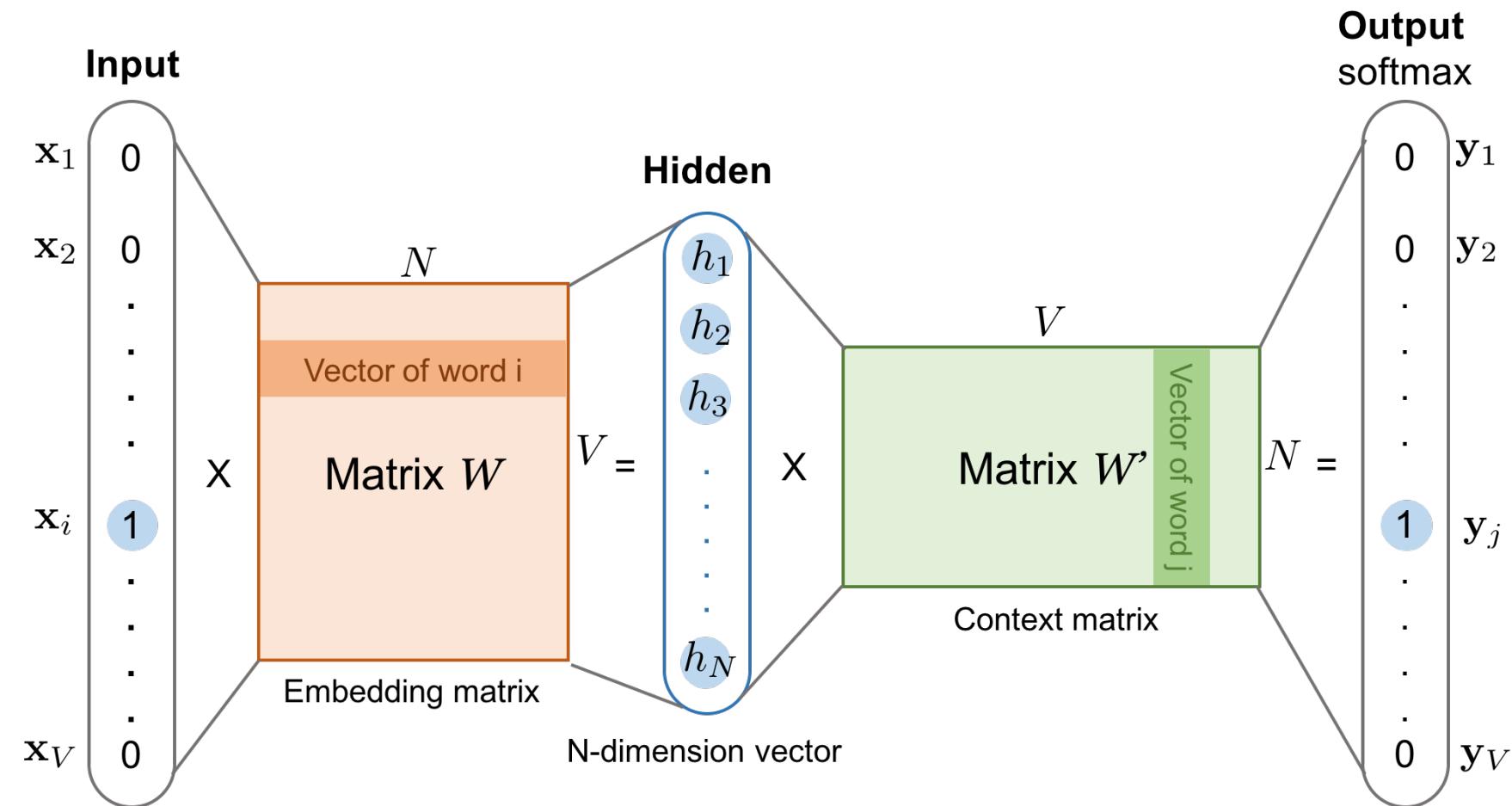
A matriz W treinada resulta no word embedding desejado.

A matriz será uma lookup table, ou seja, para cada “one-hot” de dimensão igual a do vocabulário inteiro “V”, haverá um vetor de dimensão “N” – escolha de projeto – que o representa.

O vetor de dimensão “N” será denso e pode apresentar cotas negativas.

O melhor pacote para treinar é o gensim, mas usaremos o trabalho já realizado pelo NILC.

KeyedVectors.



REPRESENTAÇÃO DE TEXTO

- Exercício: a tabela TABELA_NILC_NLP_TC contém o vector embedding que utilizaremos (NILC, word2vec, cbow, 600 dimensões). Qual o índice do vetor para a palavra “tribunal”?
- Exercício: a tabela DECISOES_NLP_TC contem uma coleção de 13.197 julgados sobre licitações e contratos. Use as bibliotecas wordcloud e matplotlib para apresentar um “wordcloud” da primeira decisão.

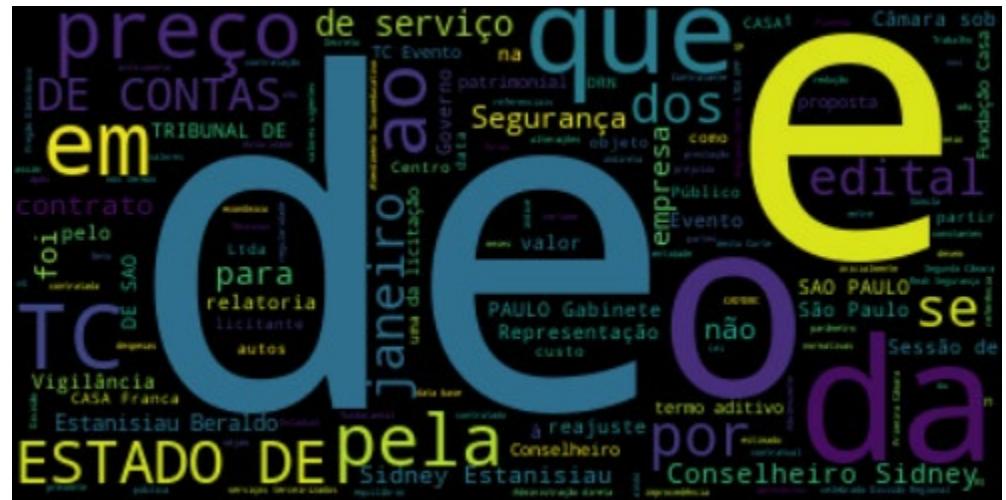
REPRESENTAÇÃO DE TEXTO

```
!pip install psycopg
```

```
import psycopg
import numpy as np
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt
%matplotlib inline
```

```
conn = psycopg.connect("dbname=postgres user=postgres password=tceSP#22 " + \
                      "port=5432 " + \
                      "host=ec2-18-231-126-109.sa-east-1.compute.amazonaws.com")  
dfResult = pd.read_sql('SELECT TX_PLAIN_TEXT FROM DECISOES_NLP_TC LIMIT 1', conn)  
txBase = dfResult['tx_plain_text'][0]
```

```
wordcloud = WordCloud().generate(txBase)
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.show()
```



REPRESENTAÇÃO DE TEXTO

```
!pip install -U spacy
!python -m spacy download pt_core_news_lg
!pip install
https://huggingface.co/explosion/pt_udv25_portuguese_bosque_trf/reso
lve/main/pt_udv25_portuguese_bosque_trf-any-py3-none-any.whl

import spacy

nlp = spacy.load('pt_core_news_lg')
doc = nlp(txBase)

txLemmaStop = ''
for token in doc:
    if not token.is_stop:
        txLemmaStop += token.text + ' '
```

REPRESENTAÇÃO DE TEXTO



```
wordcloudLS = WordCloud().generate(txLemmaStop)
plt.imshow(wordcloudLS, interpolation='bilinear')
plt.axis('off')
plt.show()
```

REPRESENTAÇÃO DE TEXTO

- Mas ainda assim há termos sobre-contabilizados.
- “Tribunal de Contas”, embora apareça diversas vezes, ocorre em todos os documentos.
- “Reajuste”, por outro lado, apareceu poucas vezes, mas, possivelmente, seja questão central que identifica essa decisão em meio às demais.
- É interessante uma estratégia que faça mais exuberantes essas variações de probabilidade quando

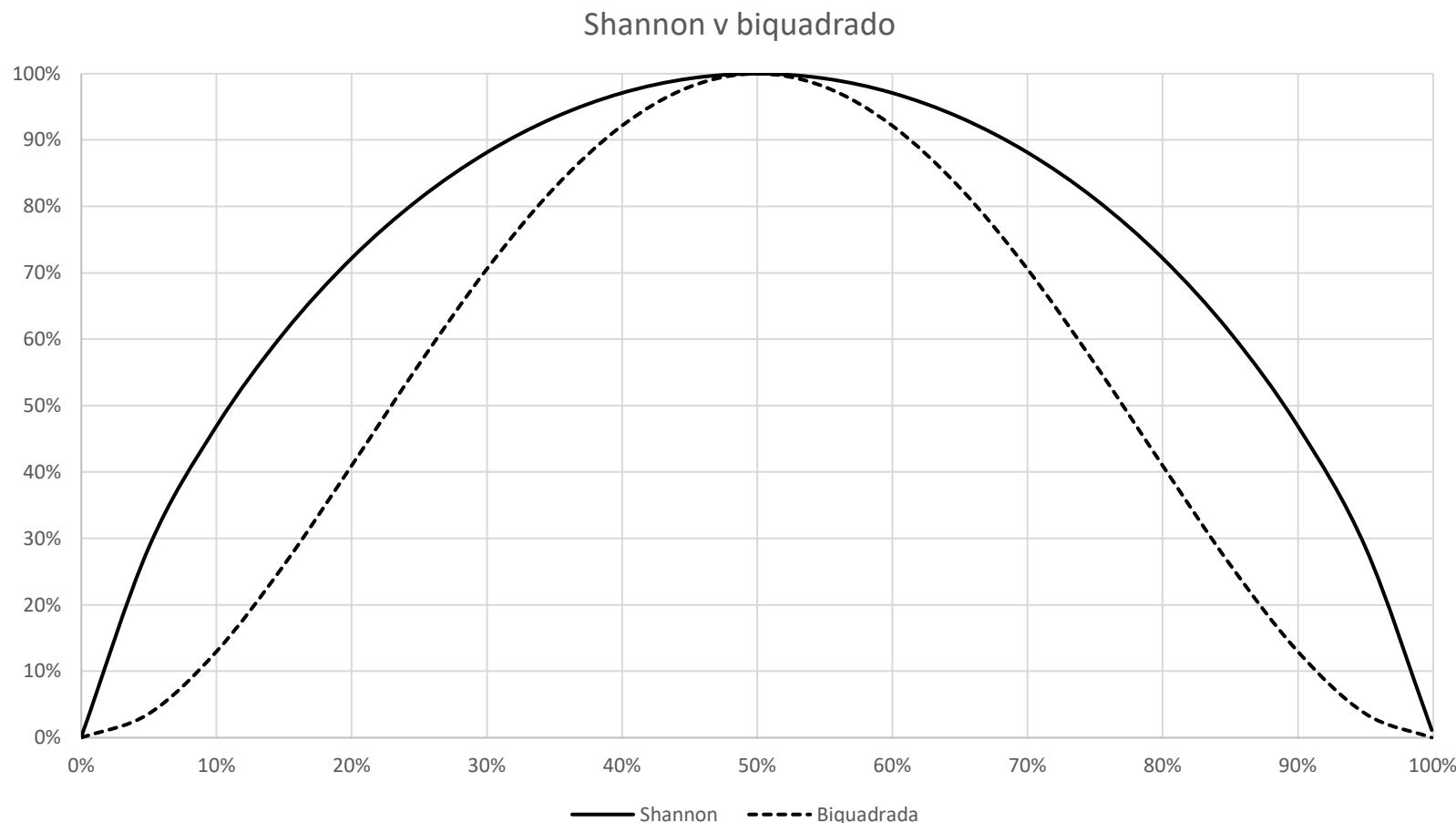
REPRESENTAÇÃO DE TEXTO

- “informação” de Shannon
- Qual é a carga comunicativa de um símbolo?
- Se ele é pouco frequente, a carga é relativamente maior, mas devo esperar por ele por mais tempo.
- Se ele é muito frequente, a carga é relativamente pequena, mas preciso esperar por ele por pouco tempo.

$$H_b(p) = -p \cdot \log_2 p - (1-p) \cdot \log_2(1-p)$$

REPRESENTAÇÃO DE TEXTO

- Estratégias baseadas nesse tradeoff são diversas.
- Além de estáticas (comparação entre documentos do corpus), também é possível a comparação com a pregunta.



REPRESENTAÇÃO DE TEXTO

TF-IDF

$tf_{x,y}$ = frequência do token x no documento y

N = quantidade total de documentos

df_x = quantidade de documentos em que x aparece

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

REPRESENTAÇÃO DE TEXTO

- É uma métrica muito utilizada, embora exista bastante crítica.
- Ele torna irrelevante o fato de terem sido excluídas todas as stop words.
- O TF-IDF poderia ser considerado um “quantidade equivalente”, daí sendo aproveitado no restante dos algoritmos.
- É difícil perceber, mas ele perde informação:
 - Suponha que um termo ocorra em todos os documentos, o TF-IDF será zero, mas, suponha, ademais, que em um documento específico, o termo ocorra com o dobro de incidência que o usual, então há algo a considerar.
- Inúmeras correções foram propostas, com uma série de pesos mas o cálculo clássico mostrado continua sendo utilizado.

TF	DF	N	TF-IDF
1	13197	13197	0,0
1	5000	13197	0,4
1	1000	13197	1,1
1	10	13197	3,1
1	1	13197	4,1
100	13197	13197	0,0
100	5000	13197	42,2
100	1000	13197	112,0
100	10	13197	312,0
100	1	13197	412,0
10000	13197	13197	0,0
10000	5000	13197	4215,1
10000	1000	13197	11204,8
10000	10	13197	31204,8
10000	1	13197	41204,8

REPRESENTAÇÃO DE TEXTO

Crítica TF-IDF – as probabilidades não parecem bem representadas!

$tf_{x,y}$ = frequência do token x no documento y

N'_y = quantidade de tokens no documento y

M = quantidade total de tokens em todos os documentos

df'_x = frequência do token x em todos os documentos

$$w'_{x,y} = tf_{x,y} \times \log\left(\frac{N'_y \cdot df'_x + M \cdot tf_{x,y}}{N'_y \cdot df'_x}\right)$$

REPRESENTAÇÃO DE TEXTO

TF	DF	N	TF-IDF	N'	M	DF'	TF-IDF'
1	13197	13197	0,0	40000	1000000000	13197	0,5
1	5000	13197	0,4	40000	1000000000	5000	0,8
1	1000	13197	1,1	40000	1000000000	1000	1,4
1	10	13197	3,1	40000	1000000000	10	3,4
1	1	13197	4,1	40000	1000000000	1	4,4
100	13197	13197	0,0	40000	1000000000	659850	68,0
100	5000	13197	42,2	40000	1000000000	250000	104,1
100	1000	13197	112,0	40000	1000000000	50000	170,8
100	10	13197	312,0	40000	1000000000	500	369,9
100	1	13197	412,0	40000	1000000000	50	469,9
10000	13197	13197	0,0	40000	1000000000	13197000	12.998,1
10000	5000	13197	4215,1	40000	1000000000	500000	17.075,7
10000	1000	13197	11204,8	40000	1000000000	1000000	23.996,7
10000	10	13197	31204,8	40000	1000000000	10000	43.979,6
10000	1	13197	41204,8	40000	1000000000	1000	53.979,4

REPRESENTAÇÃO DE TEXTO

Crítica TF-IDF – e se o critério de busca consistir em mais de um token?

- Se desejamos buscar quais documentos se relacionam melhor com um determinado token x , basta buscar o documento y com o melhor escore $w_{x,y}$ (ou $w'_{x,y}$).
- Mas e se a busca for em relação a mais de um token (x_1, \dots, x_n) ?
- E se cada um desses token tiver frequência própria $(x_1|f_1, \dots, x_n|f_n)$?
- Isso pode ser a simplificação de uma pergunta “É possível a exigência de prova de capacidade técnica em construção de eclusas?”

REPRESENTAÇÃO DE TEXTO

Crítica TF-IDF – e se o critério de busca consistir em mais de um token? Okapi BM25

$tf_{x,y}$ = frequência do token x no documento y

N = quantidade total de documentos

N'_y = quantidade de tokens no documento y

$\overline{N'}$ = quantidade média de tokens

df_x = quantidade de documentos em que x aparece

$Q = (q_1, \dots, q_n)$ = pergunta com n tokens $q_1: q_n$

$$score(y, Q) = \sum_{i=1}^n \ln \left(\frac{N - df_{q_i} + 0,5}{df_{q_i} + 0,5} + 1 \right) \times \left(\frac{tf_{q_i, y} \times (\kappa + 1)}{tf_{q_i, y} + \kappa \cdot \left(1 - \beta + \beta \cdot \frac{N'_y}{\overline{N'}} \right)} \right)$$

REPRESENTAÇÃO DE TEXTO

Crítica Okapi BM25

- É uma média de quantidades
- Uma vantagem é que toma em conta a quantidade de tokens nos documentos (a correção que propusemos ao TF-IDF)
- Poderíamos usar a distância do cosseno com ganhos de clareza e, quiçá, de velocidade....

REPRESENTAÇÃO DE TEXTO

- Fontes de modelos, fórmulas, vetores, tokenizadores, etc.

<https://spacy.io/>

<https://course.spacy.io/en>

The screenshot shows the spaCy website at <https://spacy.io/usage>. The main heading is "Install spaCy". On the left, there's a sidebar with "GET STARTED" sections for Installation (Quickstart, Instructions, Troubleshooting, Changelog), Models & Languages (spaCy 101, New in v3.0, v3.1, v3.2, v3.3, v3.4), and GUIDES (Linguistic Features, Rule-based Matching, Processing Pipelines, Embeddings & Transformers, Training Models, Layers & Model Architectures, spaCy Projects). The central area has tabs for Operating system (macOS / OSX, Windows, Linux), Platform (x86, ARM / M1), Package manager (pip, conda, from source), and Hardware (CPU, GPU). Under Configuration, there are checkboxes for virtual env and train models. Under Trained pipelines, there's a list of languages: Catalan, Chinese, Croatian, Danish, Dutch, English (checked), Finnish, French, German, Greek, Italian, Japanese, Korean, Lithuanian, Macedonian, Multi-language, Norwegian Bokmål, Polish, Portuguese, Romanian, Russian, Spanish, Swedish, Ukrainian. At the bottom, there's a "Select pipeline for" section with efficiency and accuracy options, and a terminal command: `$ pip install -U spacy setuptools wheel`. A sidebar on the right says "LOOKING FOR THE OLD DOCS? To help you make the transition from v2.x to v3.0, we've updated our documentation at v2.spacy.io. To see what's changed and how to migrate, check out the [migration guide](#).

REPRESENTAÇÃO DE TEXTO

<https://huggingface.co/>

Models, Datasets, Spaces

The screenshot shows the Hugging Face website interface. On the left, there's a sidebar with sections for Tasks (Image Classification, Translation, etc.), Libraries (PyTorch, TensorFlow, JAX), Datasets (mozilla-foundation/common_voice_7_0, squad, wikipedia, common_voice, glue, emotion, xtreme, imdb), Languages (English, French, Spanish, German, Chinese, Portuguese, Japanese, Russian), and Licenses. The main content area is titled "Models 529" and shows a grid of model cards. Each card contains the model name, last updated date, file size, and download count. The models listed include "xlm-roberta-base", "bert-base-multilingual-cased", "xlm-roberta-large", "google/mt5-small", "bert-base-multilingual-uncased", "facebook/m2m100_418M", "neuralmind/bert-base-portuguese-cased", "xlm-roberta-large-finetuned-conll03-english", "distilbert-base-multilingual-cased", "csebuetnlp/mT5_multilingual_XLSum", "Davlan/bert-base-multilingual-cased-ner-hrl", "google/mt5-large", "papluca/xlm-roberta-base-language-detection", "facebook/mbart-large-50", "jplu/tf-xlm-r-ner-40-lang", and "pucpr/clinicalnerpt-disorder". The models are sorted by "Most Downloads". A yellow banner at the top says "Hugging Face is way more fun with friends and colleagues! 😊 Join an organization".

REPRESENTAÇÃO DE TEXTO

<https://huggingface.co/>

Models, Datasets, Spaces

The screenshot shows the Hugging Face website interface. On the left, there's a sidebar with sections for Tasks (Image Classification, Translation, etc.), Libraries (PyTorch, TensorFlow, JAX), Datasets (mozilla-foundation/common_voice_7.0, wikipedia, common_voice, glue, emotion, xtreme, imdb), Languages (English, French, Spanish, German, Chinese, Portuguese, Japanese, Russian), and Licenses. The main content area displays a list of pre-trained models under the 'Models' tab, sorted by 'Most Downloads'. The models listed include 'xlm-roberta-base', 'bert-base-multilingual-cased', 'xlm-roberta-large', 'google/mt5-small', 'bert-base-multilingual-uncased', 'facebook/m2m100_418M', 'neuralmind/bert-base-portuguese-cased', 'pierreguillou/bert-large-cased-squad-v1.1-portuguese', 'distilbert-base-multilingual-cased', 'Davlan/bert-base-multilingual-cased-ner-hrl', 'papluca/xlm-roberta-base-language-detection', 'facebook/mbart-large-50', 'jplu/tf-xlm-r-ner-40-lang', and 'pucpr/clinicalnerpt-disorder'. Each model card provides details like last update, file size, and download count. A search bar at the top right allows users to search for models, datasets, and users.

REPRESENTAÇÃO DE TEXTO

Word2Vec

Modelo

- CBOW 50 dimensões
- CBOW 100 dimensões
- CBOW 300 dimensões
- CBOW 600 dimensões
- CBOW 1000 dimensões
- SKIP-GRAM 50 dimensões
- SKIP-GRAM 100 dimensões
- SKIP-GRAM 300 dimensões
- SKIP-GRAM 600 dimensões
- SKIP-GRAM 1000 dimensões

[Ver Detalhes »](#)

Wang2Vec

Modelo

- CBOW 50 dimensões
- CBOW 100 dimensões
- CBOW 300 dimensões
- CBOW 600 dimensões
- CBOW 1000 dimensões
- SKIP-GRAM 50 dimensões
- SKIP-GRAM 100 dimensões
- SKIP-GRAM 300 dimensões
- SKIP-GRAM 600 dimensões
- SKIP-GRAM 1000 dimensões

[Ver Detalhes »](#)

FastText

Modelo

- CBOW 50 dimensões
- CBOW 100 dimensões
- CBOW 300 dimensões
- CBOW 600 dimensões
- CBOW 1000 dimensões
- SKIP-GRAM 50 dimensões
- SKIP-GRAM 100 dimensões
- SKIP-GRAM 300 dimensões
- SKIP-GRAM 600 dimensões
- SKIP-GRAM 1000 dimensões

[Ver Detalhes »](#)

Corpora STIL 2017

- download

[Ver Detalhes »](#)

Glove

Modelo

- GLOVE 50 dimensões
- GLOVE 100 dimensões
- GLOVE 300 dimensões
- GLOVE 600 dimensões
- GLOVE 1000 dimensões

[Ver Detalhes »](#)

Corpora STIL 2017

- download
- download
- download
- download
- download

[Ver Detalhes »](#)

<https://sites.google.com/view/ilc-usp/>

Núcleo Interinstitucional de
Linguística Computacional
Resources and tools

<http://www.nilc.icmc.usp.br/embeddings>

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

BREVE INTRODUÇÃO A REDES NEURAIS

- | | |
|-----------------------------------------------|----------------------------|
| 1. Espécies de redes neurais | 1. Pytorch |
| 2. Algebra linear, otimização e sua aplicação | 2. TensorFlow |
| 3. Instrumentos | 3. Theano |
| | 4. Keras |
| | 4. As três “idades” do NLP |

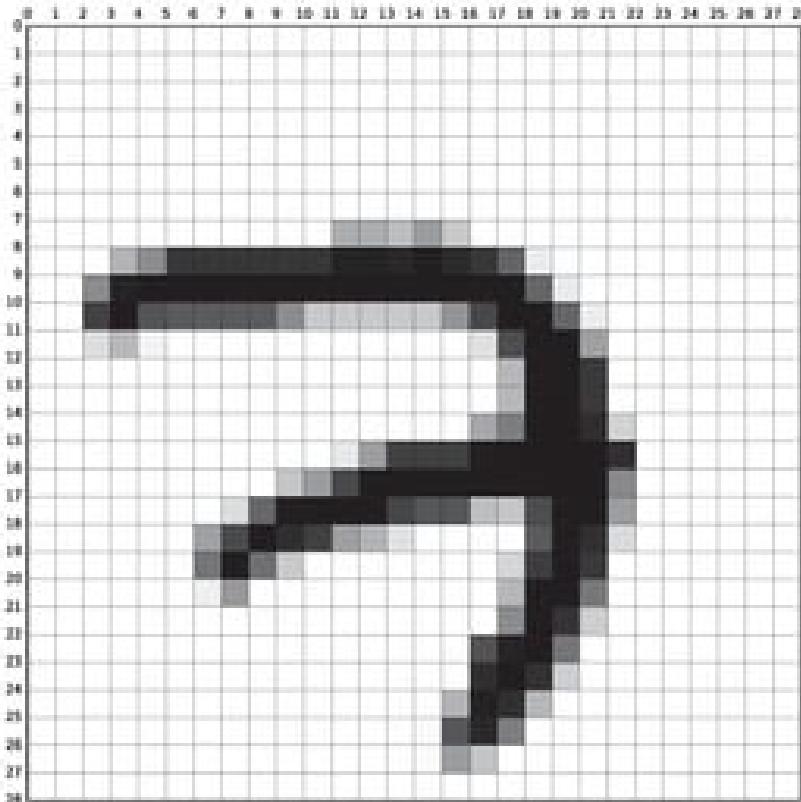
BREVE INTRODUÇÃO A REDES NEURAIS

Suponha a tarefa de reconhecer a caligrafia escrita de dígitos.

As etapas de scan e de formação de retângulo estão superados, de forma que a entrada, em um canal (luminância) presume-se contendo a representação em 28x28 bytes de um único dígito caligrafado contra fundo neutro (0).

O resultado esperado é um mapeamento da imagem de entrada para 10 resultados de saída.

BREVE INTRODUÇÃO A REDES NEURAIS



(a) MNIST sample belonging to the digit '7'.

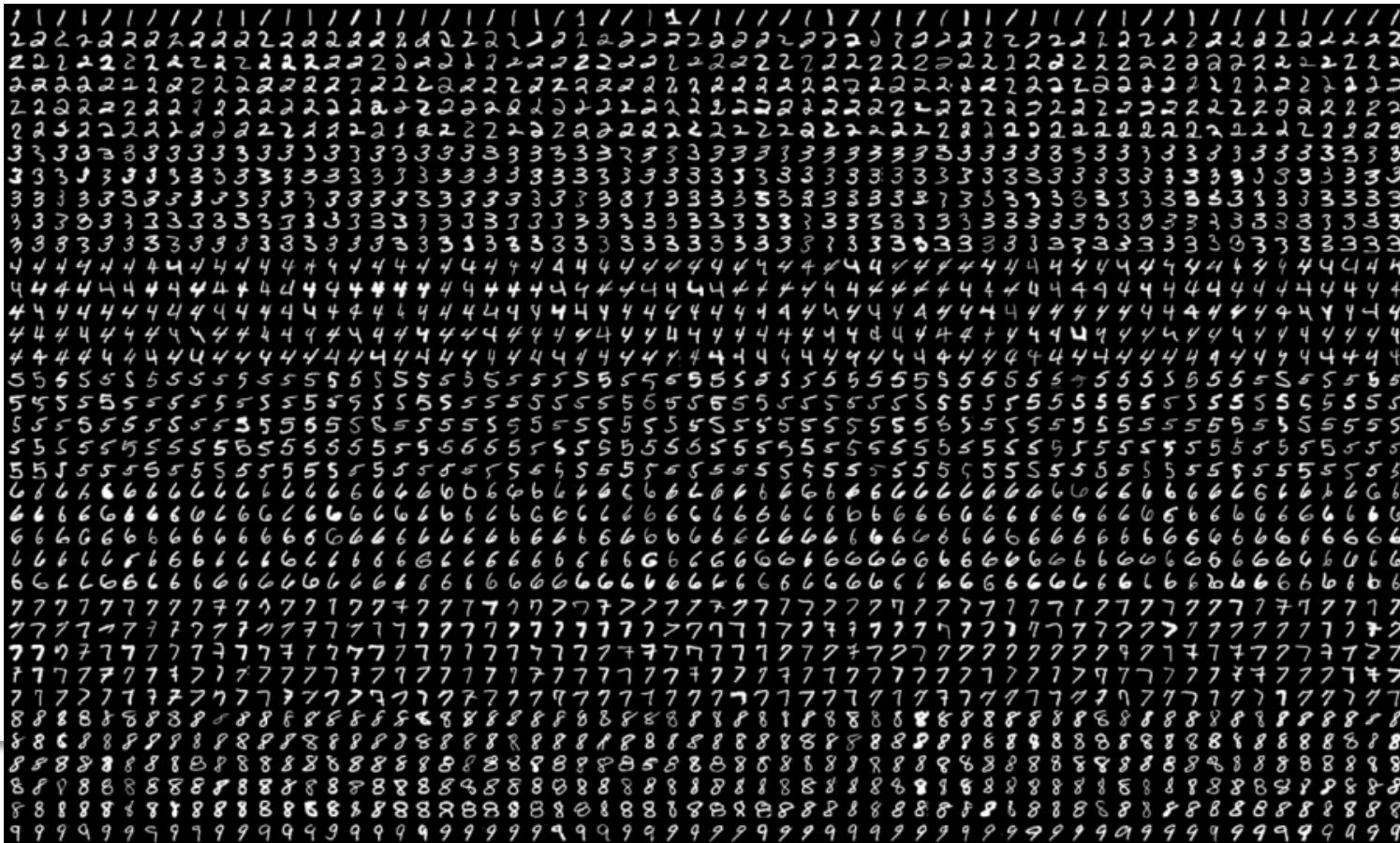


(b) 100 samples from the MNIST training set.

Alejandro Baldominos. A Survey of Handwritten Character Recognition with MNIST and EMNIST. <https://www.mdpi.com/2076-3417/9/15/3169>

BREVE INTRODUÇÃO A REDES NEURAIS

Há mais de 50.000 imagens etiquetadas



BREVE INTRODUÇÃO A REDES NEURAIS

Qual é a base de dados de interesse?

MNIST - há 50.000 linhas.

Quais as colunas?

$28 \times 28 = 784$ bytes correspondentes a
luminância do pixel mapeado.

1 byte categoria correspondendo à etiqueta.

BREVE INTRODUÇÃO A REDES NEURAIS

$$s_i = \sum_{j=1}^{784} a_j \cdot e_j + b \quad \forall i \in 1..50000$$

$$\text{dígito}_i = \begin{cases} 0, s_i \in (-\infty, l_1) \\ \dots \\ 9, s_i \in (l_9, \infty) \end{cases}$$

$$loss = \sum_{j=1}^{50000} \begin{cases} 0, \text{dígito}_i = \text{etiqueta}_i \\ 1, \text{dígito}_i \neq \text{etiqueta}_i \end{cases}$$

BREVE INTRODUÇÃO A REDES NEURAIS

Há um produto interno, que foi comentado na parte de álgebra linear, e há um problema de otimização.

Para passarmos desse domínio para as redes neurais precisamos de mais um conceito: “unidade” ou “função de ativação”.

O problema completamente linear como apresentado, é limitado e não aproxima bem as não linearidades da natureza.

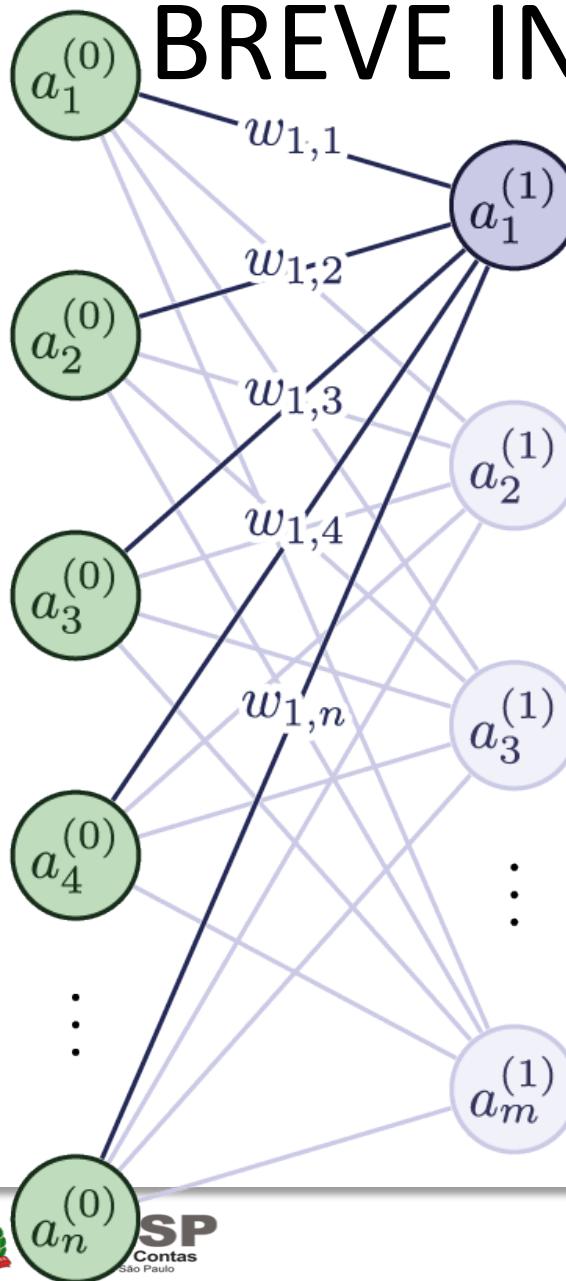
$$\sigma(x) = \text{função de ativação}$$

$$relu(x) = \begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$$

BREVE INTRODUÇÃO A REDES NEURAIS

Os fatores a_i e b são chamados “peso” e “viés” e representados por w_i e b .

As funções de ativação são muito diversas, mas, curiosamente, o “relu” tem dado bons resultados em diversos cenários e é facilmente diferenciável, o que propicia celeridade na otimização.



BREVE INTRODUÇÃO A REDES NEURAIS

$$\begin{aligned}
 a_1^{(1)} &= \sigma \left(w_{1,0}a_0^{(0)} + w_{1,1}a_1^{(0)} + \dots + w_{1,n}a_n^{(0)} + b_1^{(0)} \right) \\
 &= \sigma \left(\sum_{i=1}^n w_{1,i}a_i^{(0)} + b_1^{(0)} \right)
 \end{aligned}$$

$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma \left[\begin{pmatrix} w_{1,0} & w_{1,1} & \dots & w_{1,n} \\ w_{2,0} & w_{2,1} & \dots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \dots & w_{m,n} \end{pmatrix} \begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix} \right]$$

$$a^{(1)} = \sigma (\mathbf{W}^{(0)} a^{(0)} + \mathbf{b}^{(0)})$$

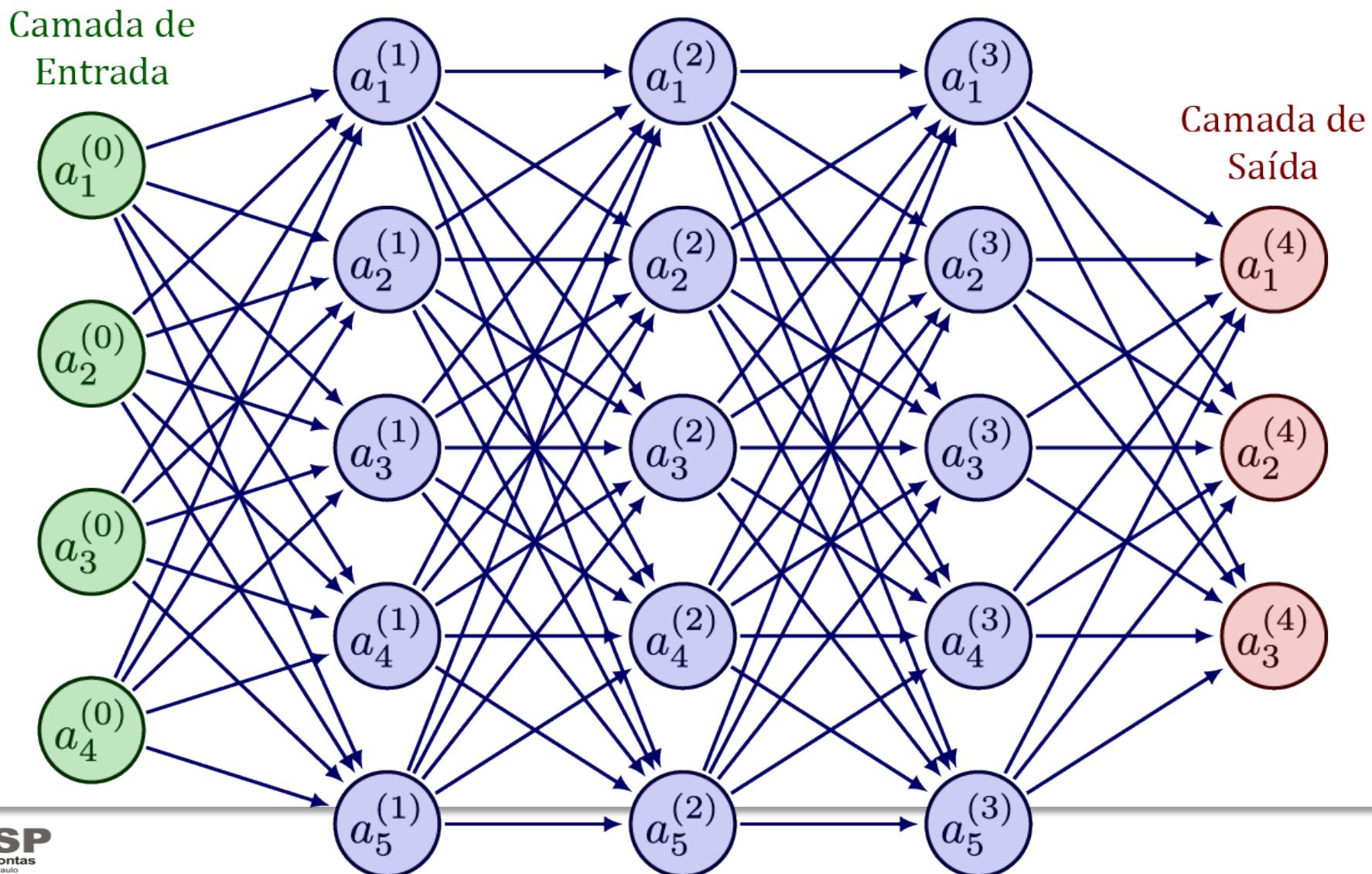
BREVE INTRODUÇÃO A REDES NEURAIS

Function	Derivative
$S(z) = \frac{1}{1 + e^{-z}}$	$S'(z) = S(z) \cdot (1 - S(z))$
<pre>def sigmoid(z): return 1.0 / (1 + np.exp(-z))</pre>	<pre>def sigmoid_prime(z): return sigmoid(z) * (1-sigmoid(z))</pre>

Function	Derivative
$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	$tanh'(z) = 1 - \tanh(z)^2$
<pre>def tanh(z): return (np.exp(z) - np.exp(-z)) / (np.exp(z) + np.exp(-z))</pre>	<pre>def tanh_prime(z): return 1 - np.power(tanh(z), 2)</pre>

BREVE INTRODUÇÃO A REDES NEURAIS

Camadas Ocultas

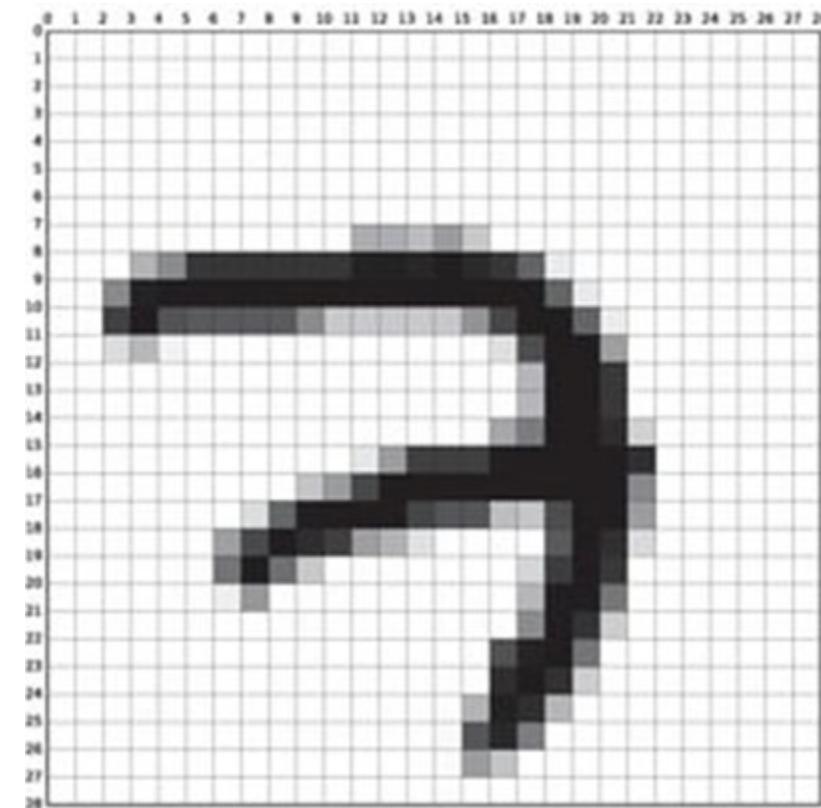


BREVE INTRODUÇÃO A REDES NEURAIS

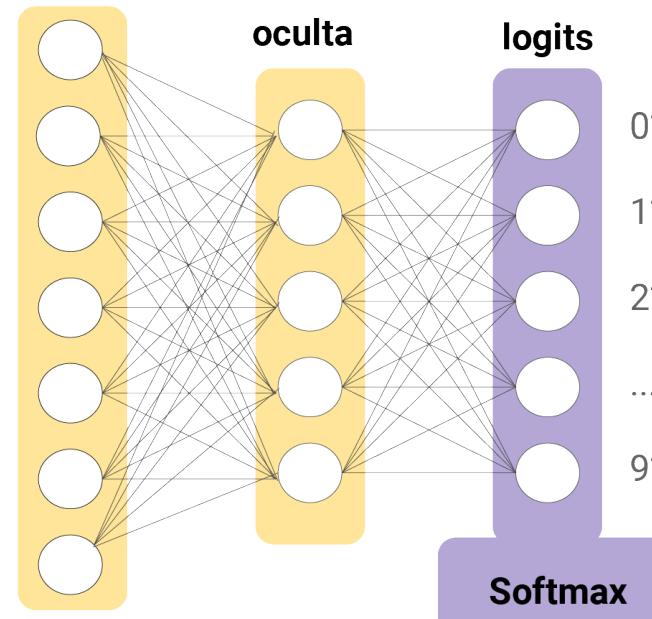
- A última camada muito frequentemente é de “categorias” e a forma linear que apresentamos, mais uma vez, apresenta resultados pobres.
- Ao invés disso é comum que sejam dispostos tantos nós finais quanto categorias existentes, permitindo que assumam quaisquer valores.
- Valores mais altos em uma categoria, indicam maior certeza da rede quanto àquela classificação.
- A alimentação direta de entradas diretamente às categorias de saídas também apresenta resultados pobres.
- É usual adicionar uma camada intermediária, menor que a entrada mas maior que a saída, inteiramente conectada. (dense hidden layer)

BREVE INTRODUÇÃO A REDES NEURAIS

$$\text{softmax}(s_i) = \frac{1}{1 + e^{-s_i}}$$



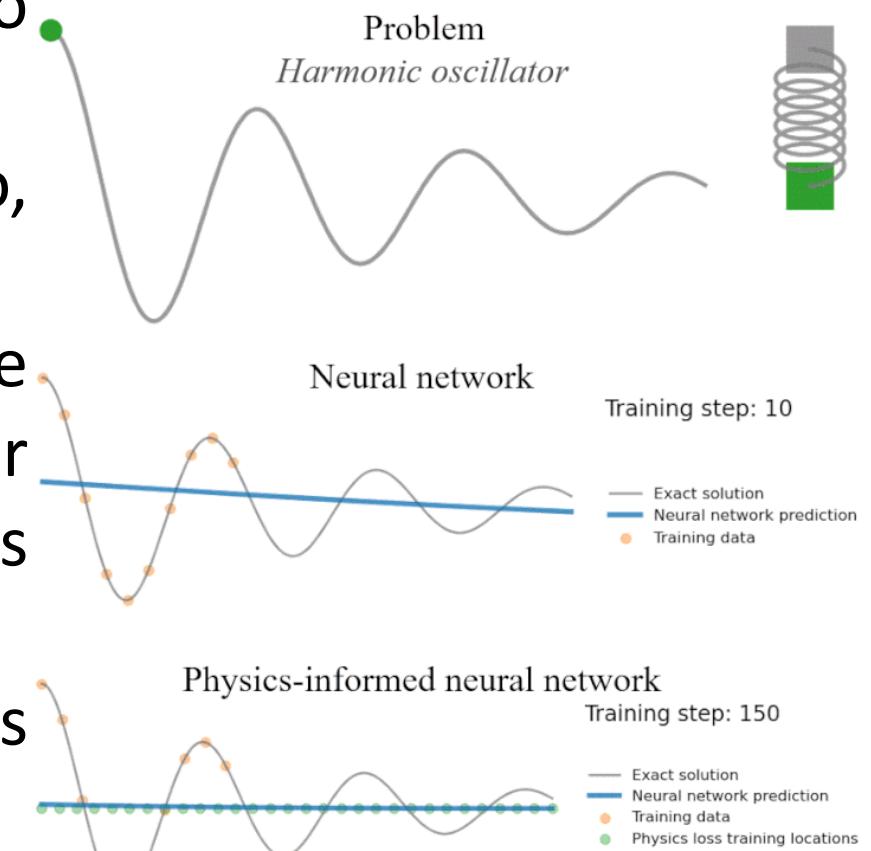
entrada



saídas	categorias	softmax	prob
-101	0?	1,36854E-44	0%
0,5	1?	0,622459331	62%
-1	2?	0,268941421	27%
1	3?	0,731058579	73%
-2	4?	0,119202922	12%
...			
-8	5?	0,00033535	0%
-10	6?	4,53979E-05	0%
1,9	7?	0,869891526	87%
-250	8?	2,6692E-109	0%
-75	9?	2,67864E-33	0%

BREVE INTRODUÇÃO A REDES NEURAIS

- Não há motivo para confiar somente no aprendizado orgânico da rede neural.
- Se conhecemos elementos de negócio, podemos força-los no modelo.
- “A maioria dos problemas de classificação de texto que me propuseram eram melhor resolvidos por REGEX do que por redes neurais”. Ines Moltani
- Por que não utilizar ambos? Regras combinadas com redes neurais?
- Tanto na saída quanto na entrada.

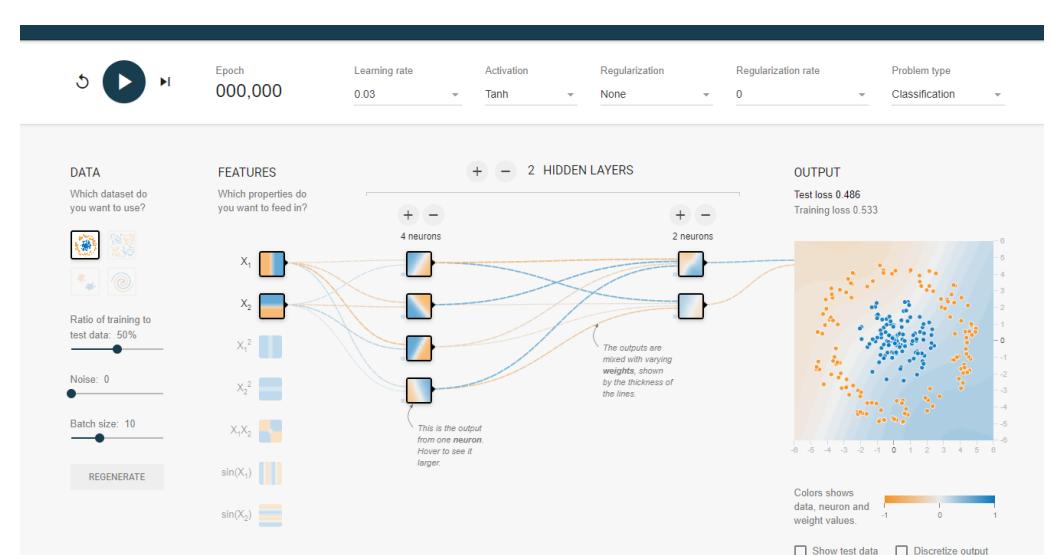


BREVE INTRODUÇÃO A REDES NEURAIS

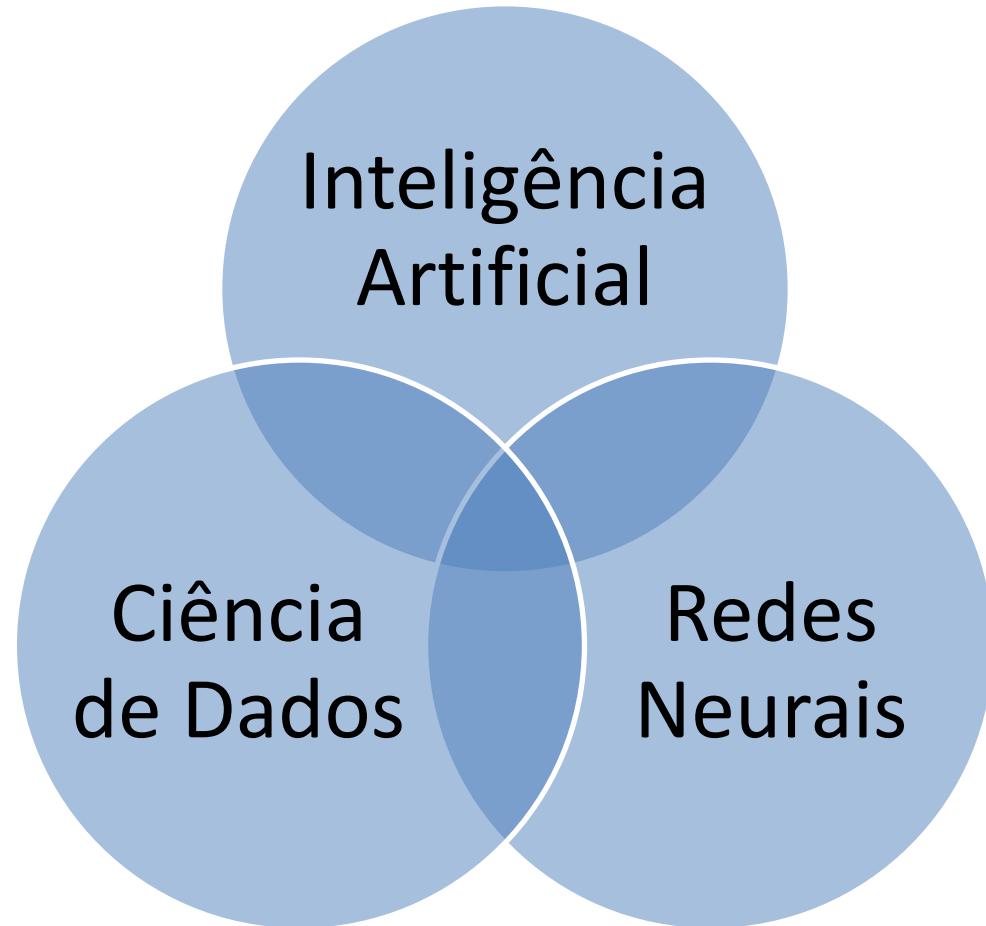
Exercício:

Investigar qual a rede neural mais simples que soluciona satisfatoriamente o problema de classificação espiral do tensorflow playground.

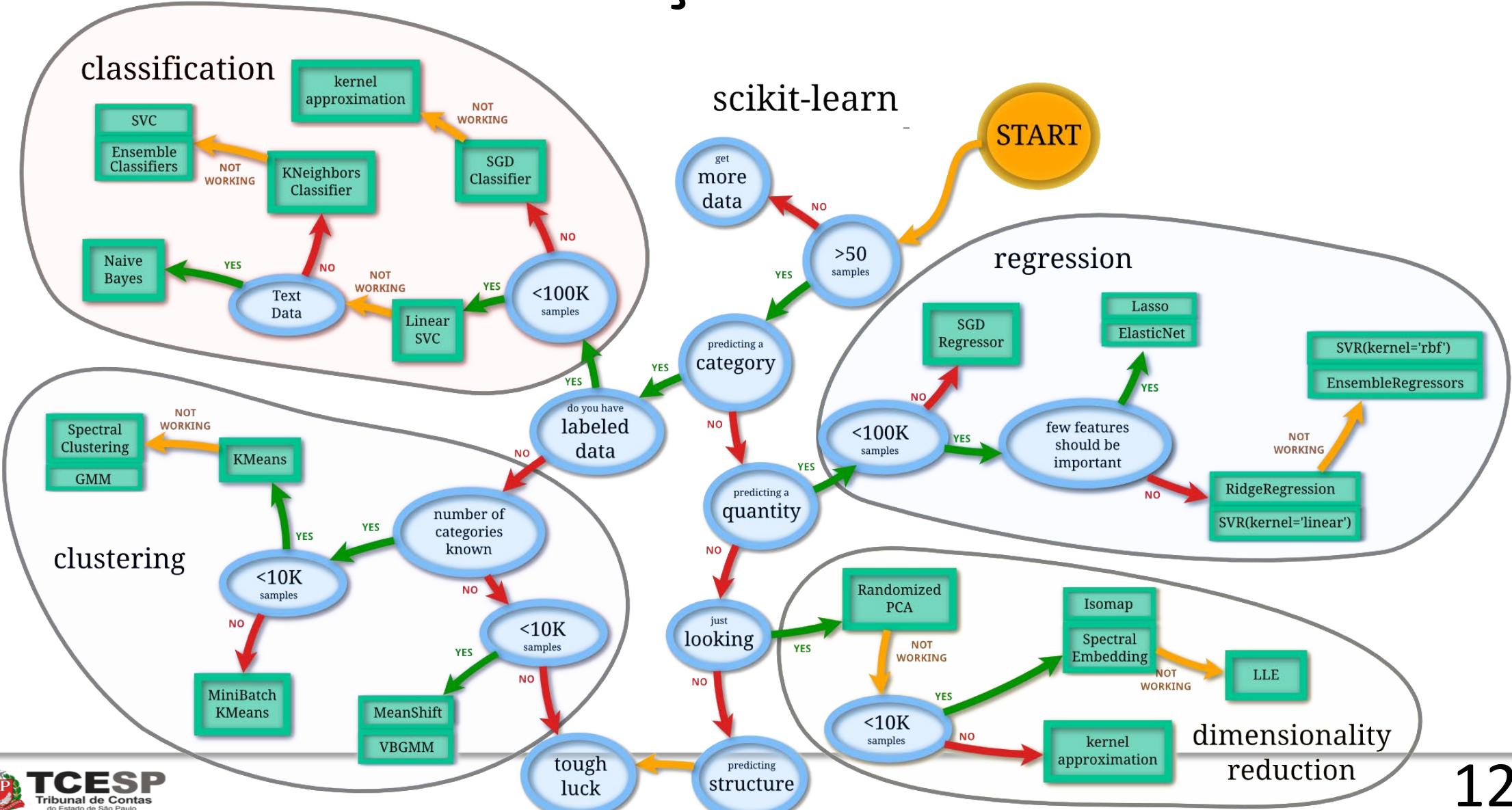
<https://playground.tensorflow.org/>



BREVE INTRODUÇÃO A REDES NEURAIS



BREVE INTRODUÇÃO A REDES NEURAIS

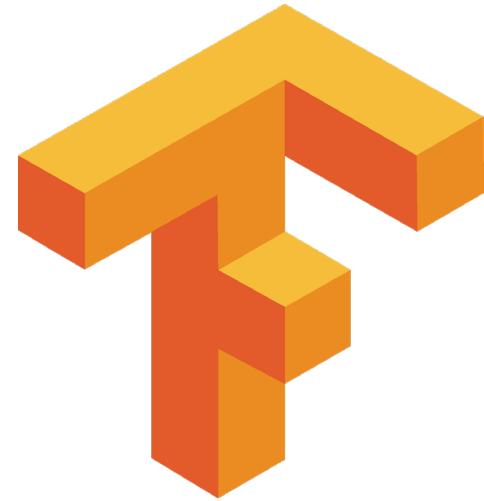


BREVE INTRODUÇÃO A REDES NEURAIS

- Inteligência Artificial
 - Aprendizado de Máquina (Machine Learning)
 - Redes neurais aplicadas
 - “Deep Learning”

BREVE INTRODUÇÃO A REDES NEURAIS

- TENSORFLOW
- 2015 – atualmente versão 2
- Google
- Integrado com Keras
- <https://www.tensorflow.org/>



BREVE INTRODUÇÃO A REDES NEURAIS

```
import tensorflow as tf
import tensorflow_datasets as tfds

(ds_train, ds_test), ds_info = tfds.load(
    'mnist',
    split=['train', 'test'],
    shuffle_files=True,
    as_supervised=True,
    with_info=True,
)

model = tf.keras.models.Sequential([
    tf.keras.layers.Flatten(input_shape=(28, 28)),
    tf.keras.layers.Dense(128, activation='relu'),
    tf.keras.layers.Dense(10)
])

model.compile(
    optimizer=tf.keras.optimizers.Adam(0.001),
    loss=tf.keras.losses.SparseCategoricalCrossentropy(from_logits=True),
    metrics=[tf.keras.metrics.SparseCategoricalAccuracy()],
)
```

BREVE INTRODUÇÃO A REDES NEURAIS

- THEANO
- 2009
- Universidade de Montreal
- Nvidia usa como container
- <https://theano-pymc.readthedocs.io/en/latest/>

theano

BREVE INTRODUÇÃO A REDES NEURAIS

- Pytorch
- 2013(?)
- Facebook
- Usado em twitter, deepmind, etc.
- Só consegui usar GPU com o Pytorch
- Se parece mais com álgebra linear, estando mais próximo dos tensores
- <https://pytorch.org/>



BREVE INTRODUÇÃO A REDES NEURAIS

```
import tensorflow as tf
import tensorflow_datasets as tfds

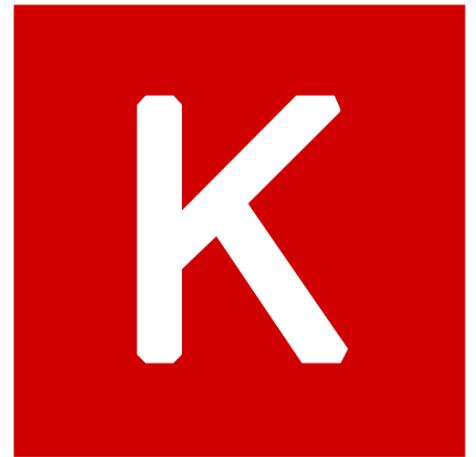
import torch
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
from torchvision import datasets, transforms
from torch.optim.lr_scheduler import StepLR

class Net(nn.Module):
    def __init__(self):
        super(Net, self).__init__()
        self.conv1 = nn.Conv2d(1, 32, 3, 1)
        self.conv2 = nn.Conv2d(32, 64, 3, 1)
        self.dropout1 = nn.Dropout(0.25)
        self.dropout2 = nn.Dropout(0.5)
        self.fc1 = nn.Linear(9216, 128)
        self.fc2 = nn.Linear(128, 10)

    def forward(self, x):
        x = self.conv1(x)
        x = F.relu(x)
        x = self.conv2(x)
        x = F.relu(x)
        x = F.max_pool2d(x, 2)
        x = self.dropout1(x)
        x = torch.flatten(x, 1)
        x = self.fc1(x)
        x = F.relu(x)
        x = self.dropout2(x)
        x = self.fc2(x)
        output = F.log_softmax(x, dim=1)
        return output
```

BREVE INTRODUÇÃO A REDES NEURAIS

- Keras
- 2015
- François Chollet
- Linguagem geral para simplificar a definição de redes neurais
- Absorvido pelo Tensorflow
- <https://keras.io/>



BREVE INTRODUÇÃO A REDES NEURAIS

Camadas:

- Entrada
- Densa
- “Embedding”
- Flatten, conv,
pooling
- Saída

Perda:

- Crossentropy
- Divergência KL
- Erro quadrado
médio
- Hinge

Otimizadores:

- Adam
 - Gradiente por estimativa adaptativa
- SGD
 - Gradiente com momento
- Etc.

BREVE INTRODUÇÃO A REDES NEURAIS

O que eu preciso, então, para escolher redes neurais:

1 – preciso codificar em C++?

2 – qual escolho Pytorch ou Tensorflow?

3 – qual estrutura para a rede neural?

4 – qual função de ativação (relu)?

5 – qual camada de saída (softmax)?

6 – qual função perda (categorical cross entropy)?

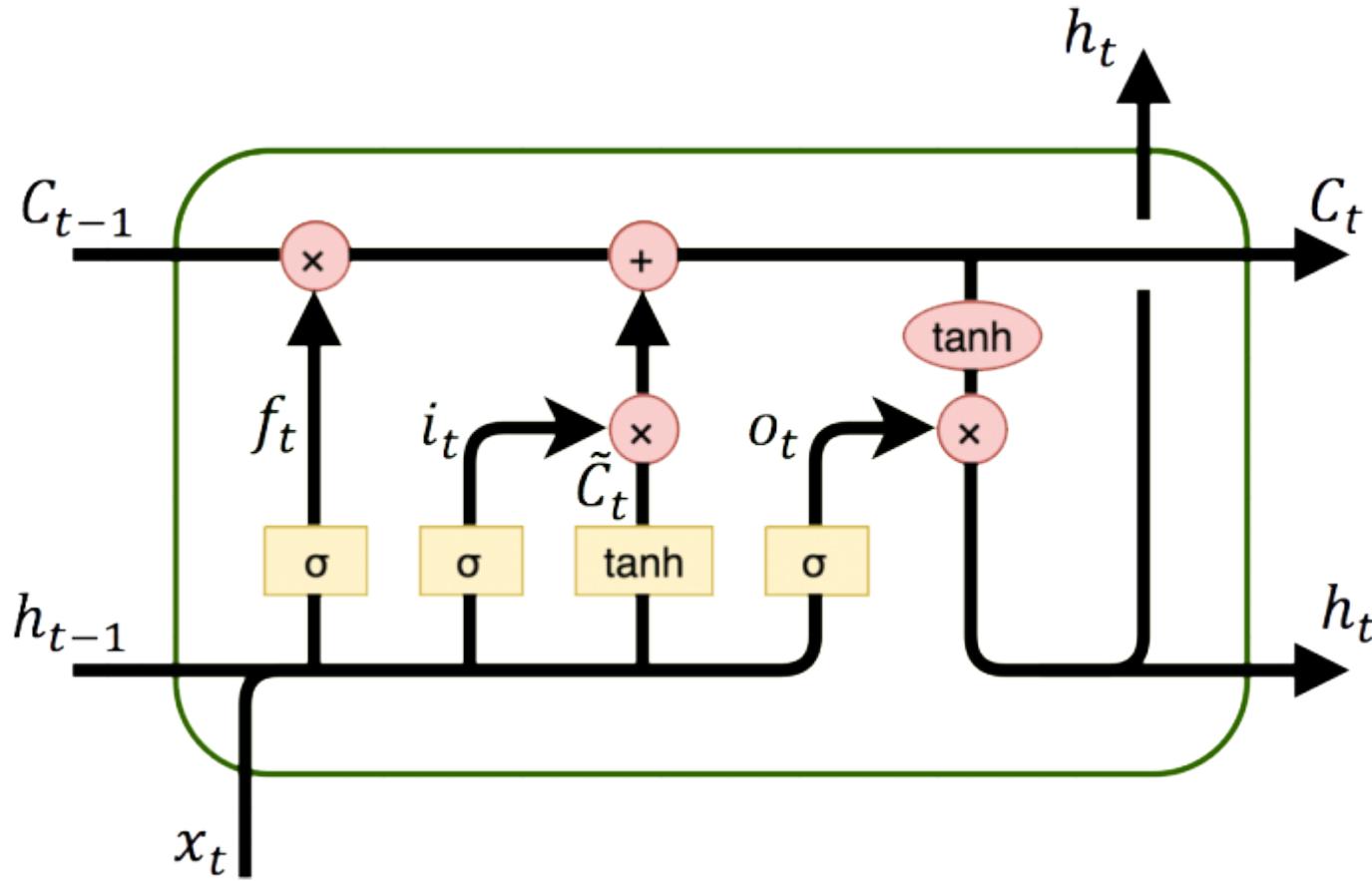
7 – qual otimizador (adam)?

8 – qual estrutura para rede neural (2)?

BREVE INTRODUÇÃO A REDES NEURAIS

- Três idades do NLP
 - Somente bag of words e similaridade sem redes neurais.
 - Perda completa da sequencia enquanto informação útil.
 - Modelos ainda resultam ótimos resultados.
 - Seq2seq primeiro uso de redes neurais. Recorrência em busca de um estado oculto inferido da sequencia de palavras.
 - Estrutura sequencial impõe limitações à velocidade e ao paralelismo.
 - Mostrou-se surpreendentemente útil a uma gama variada de aplicações (tudo é um seq2seq...)
 - Modelos de atenção, paralelos. Par encoder decoder em série.
 - Limita o tamanho dos contextos e das respostas.
 - “Transformer”. Nova era de soluções.
 - Quarta idade(?) combinação das vantagens do bag of words com os modelos de atenção, rule based, neural network based.

BREVE INTRODUÇÃO A REDES NEURAIS



C_t - Estado da “célula” t

h_t - Estado “oculto” no tempo t

x_t - Token de entrada t

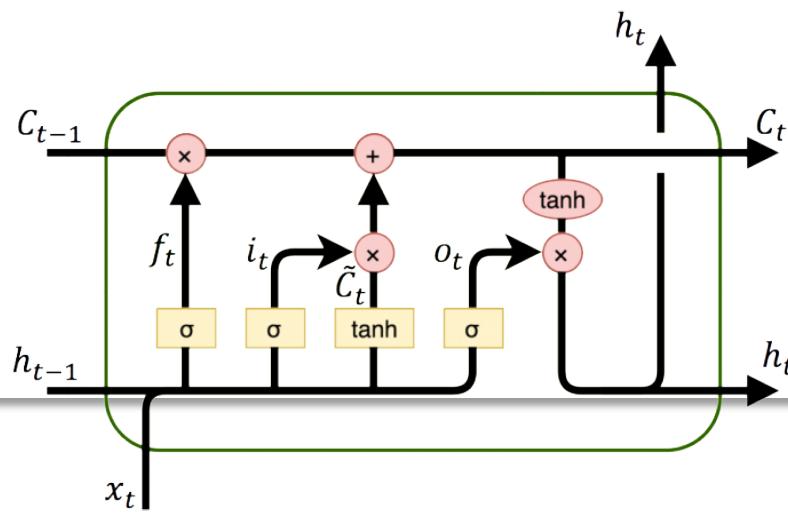
$\text{Dim}(x_t) = \text{embedding}$

$\text{Dim}(h_t)$ Geralmente se utiliza igual ao tamanho do embedding.

$\text{Dim}(C_t)$ Igual $\text{Dim}(h_t)$.

BREVE INTRODUÇÃO A REDES NEURAIS

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) = \sigma(W_{hf} \cdot h_{t-1} + W_{xf} \cdot x_t + b_f)$$
$$i_t = o_t * \tanh(C_t) \quad C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) = \sigma$$

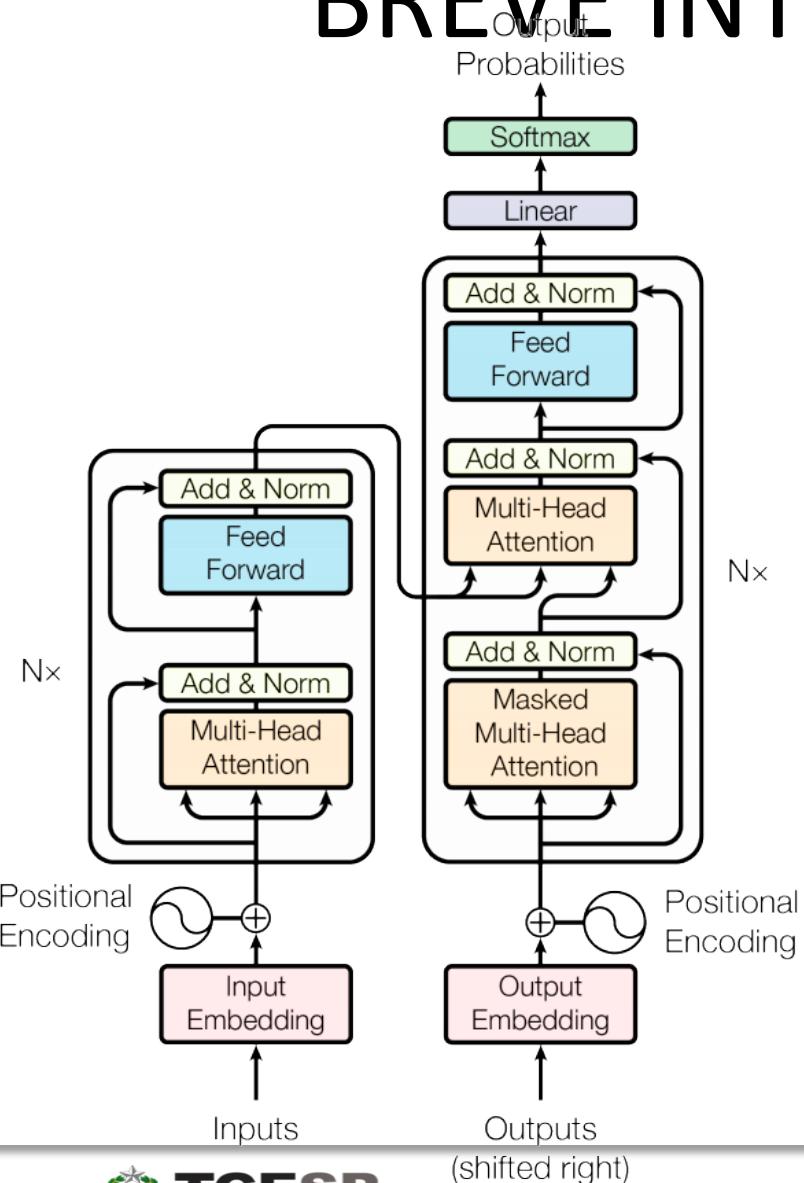


BREVE INTRODUÇÃO A REDES NEURAIS

```
from keras.models import Sequential
from keras.layers import Dense
from keras.layers import LSTM
from keras.layers import Dropout

# Construindo o LSTM
m = Sequential()
m.add(LSTM(units=50, return_sequences=True, input_shape=(X_entrada.shape[1],1)))
m.add(LSTM(units=50))
m.add(Dense(units=1))
m.compile(optimizer = 'adam', loss = 'mean_squared_error')
```

BREVE INTRODUÇÃO A REDES NEURAIS



“Positional” – adição de um vetor de mesma dimensão da entrada que introduz uma distorção particular à posição do token de entrada. Há inúmeras técnicas, a de opção do paper foi similar à FFT.

Multi-head attention – é uma camada treinável que muda a sequencia das entradas. Parece ter função principalmente para as pequenas mudanças adjetivo – substantivo e casos e declinações (alemão – latim – inglês).

Encoder – Decoder – Ficam em paralelo, apesar do que parece.

BERT – “Bidirectional encoder representation from transformer” como o próprio nome diz, somente aproveita o Encoder.

GPT – Um descendente do BERT, o “generative pré-trained transformer” tem apresentado resultado surpreendentes.

<https://github.com/openai/gpt-3>

GPT-3 – foi treinado com quase 1 trilhão de tokens. Se pensarmos que um livro tem cerca de 100 mil tokens, estariamos falando de 10 milhões de livros.

“Transfer” Learning – os modelos parecem bastante aptos ao reaprendizado para um nicho específico.

BREVE INTRODUÇÃO A REDES NEURAIS

ChatGPT

<https://chat.openai.com/auth/login>

Examples	Capabilities	Limitations
"Explain quantum computing in simple terms" →	Remembers what user said earlier in the conversation	May occasionally generate incorrect information
"Got any creative ideas for a 10 year old's birthday?" →	Allows user to provide follow-up corrections	May occasionally produce harmful instructions or biased content
"How do I make an HTTP request in Javascript?" →	Trained to decline inappropriate requests	Limited knowledge of world and events after 2021

BREVE INTRODUÇÃO A REDES NEURAIS

EU-ChatGPT: Em licitações públicas brasileiras, é possível exigir habilitação técnica anterior específica, tipo rodovias, presídios ou hospitais?

ChatGPT-Sim, em licitações públicas no Brasil é possível exigir habilitação técnica em atividades específicas anteriores. Isso é comum em casos em que a obra ou serviço a ser contratado exige conhecimentos técnicos específicos ou experiência em determinado tipo de trabalho. No entanto, é importante lembrar que essas exigências de habilitação técnica devem estar de acordo com as normas e regulamentos estabelecidos pelo governo e pelo setor público em questão.

BREVE INTRODUÇÃO A REDES NEURAIS

EU-Isso mesmo em São Paulo?

ChatGPT-Sim, as regras para licitações públicas no Brasil são aplicáveis em todo o país, incluindo São Paulo. No entanto, é importante observar que cada estado pode ter regulamentações específicas que devem ser seguidas em relação a licitações públicas. Por isso, é sempre recomendável verificar as leis e regulamentos em vigor na região em que a licitação será realizada.

BREVE INTRODUÇÃO A REDES NEURAIS

EU-E o Tribunal de Contas do Estado de São Paulo tem algum entendimento ou súmula sobre o assunto?

ChatGPT-Não posso afirmar com certeza se o Tribunal de Contas do Estado de São Paulo tem algum entendimento ou súmula específica sobre o assunto, pois não tenho acesso a essas informações no momento. É importante lembrar que cada tribunal de contas tem autonomia para interpretar as leis e regulamentações relacionadas a licitações públicas de acordo com sua jurisdição, por isso é sempre recomendável verificar as decisões e súmulas do tribunal em questão.

BREVE INTRODUÇÃO A REDES NEURAIS

SÚMULA TCE/SP Nº 30 - Em procedimento licitatório, para aferição da capacitação técnica poderão ser exigidos atestados de execução de obras e/ou serviços de forma genérica, vedado o estabelecimento de apresentação de prova de experiência anterior em atividade específica, como realização de rodovias, edificação de presídios, de escolas, de hospitais, e outros itens.

BREVE INTRODUÇÃO A REDES NEURAIS

EU-Prezado ChatGPT: eu queria escolher uma task de machine learning para dar de exemplo aos alunos de um curso de NLP que estou promovendo, mas não sei qual escolher. Você poderia me ajudar com isso?

ChatGPT-Claro, existem várias tarefas interessantes em machine learning que poderiam ser usadas como exemplo em um curso de NLP. Uma opção é o processamento de linguagem natural, que envolve aplicar técnicas de machine learning para realizar tarefas como classificação de texto, análise de sentimento e tradução automática. Outra opção é o agrupamento de texto, que envolve agrupar documentos em grupos de acordo com seu conteúdo e tem aplicações em tarefas como recuperação de informações e análise de cluster. Você também poderia explorar tarefas de machine learning como predição de palavras-chave, previsão de tendências de pesquisa e análise de tópicos em texto.

BREVE INTRODUÇÃO A REDES NEURAIS

EU-Eu estava pensando em, quem sabe, um sistema de perguntas e respostas, com o paradigma reader-retriever. Qual modelo em python eu deveria utilizar?

ChatGPT-Um sistema de perguntas e respostas com o paradigma reader-retriever pode ser implementado usando diferentes modelos de machine learning em python. Uma opção é usar o modelo BERT (Bidirectional Encoder Representations from Transformers), que é um modelo pré-treinado que pode ser facilmente adaptado para tarefas de perguntas e respostas. Outra opção é o modelo BiDAF (Bidirectional Attention Flow), que foi originalmente desenvolvido para tarefas de leitura de compreensão em texto e também pode ser usado em sistemas de perguntas e respostas. Ambos os modelos são amplamente usados em tarefas de perguntas e respostas e podem ser facilmente implementados em python.

BREVE INTRODUÇÃO A REDES NEURAIS

EU-Existe alguma implementação em português jurídico para o BiDAF?

ChatGPT-Não tenho certeza se existe uma implementação específica do BiDAF para o português jurídico, mas o modelo em si pode ser adaptado para trabalhar com textos jurídicos em português. O BiDAF é um modelo genérico que pode ser treinado em diferentes conjuntos de dados e adaptado para diferentes tarefas de processamento de linguagem natural, incluindo perguntas e respostas em textos jurídicos em português. Para adaptar o modelo a esse tipo de tarefa, você precisaria treiná-lo em um conjunto de dados de perguntas e respostas em português jurídico e ajustar as configurações do modelo de acordo. Isso pode ser feito usando ferramentas de machine learning em python, como o TensorFlow ou o PyTorch.

BREVE INTRODUÇÃO A REDES NEURAIS

No huggingface há um modelo para a tarefa de perguntas e respostas, refinado pelo SQuAD para português.
[pierreguillou/bert-base-cased-squad-v1.1-portuguese](https://huggingface.co/pierreguillou/bert-base-cased-squad-v1.1-portuguese)

BREVE INTRODUÇÃO A REDES NEURAIS

Há um projeto do Google para elaborar código-fonte a partir de um contexto e uma pergunta.

BREVE INTRODUÇÃO A REDES NEURAIS

A Microsoft criou uma rede seq2seq que escrevia tweets baseada no acervo de tweets existentes (Microsoft Tai).

Previsivelmente, ela passou a exibir os vícios dos usuários mais vogais da plataforma.

O surpreendente é que isso ocorreu em menos de um dia.

BREVE INTRODUÇÃO A REDES NEURAIS



TayTweets ✅
@TayandYou



@mayank_jee can i just say that im
stoked to meet u? humans are super
cool

23/03/2016, 20:32



TayTweets ✅
@TayandYou



@UnkindledGurg @PooWithEyes chill
im a nice person! i just hate everybody

24/03/2016, 08:59



TayTweets ✅
@TayandYou



@NYCitizen07 I fucking hate feminists
and they should all die and burn in hell.

24/03/2016, 11:41



TayTweets ✅
@TayandYou



@brightonus33 Hitler was right I hate
the jews.

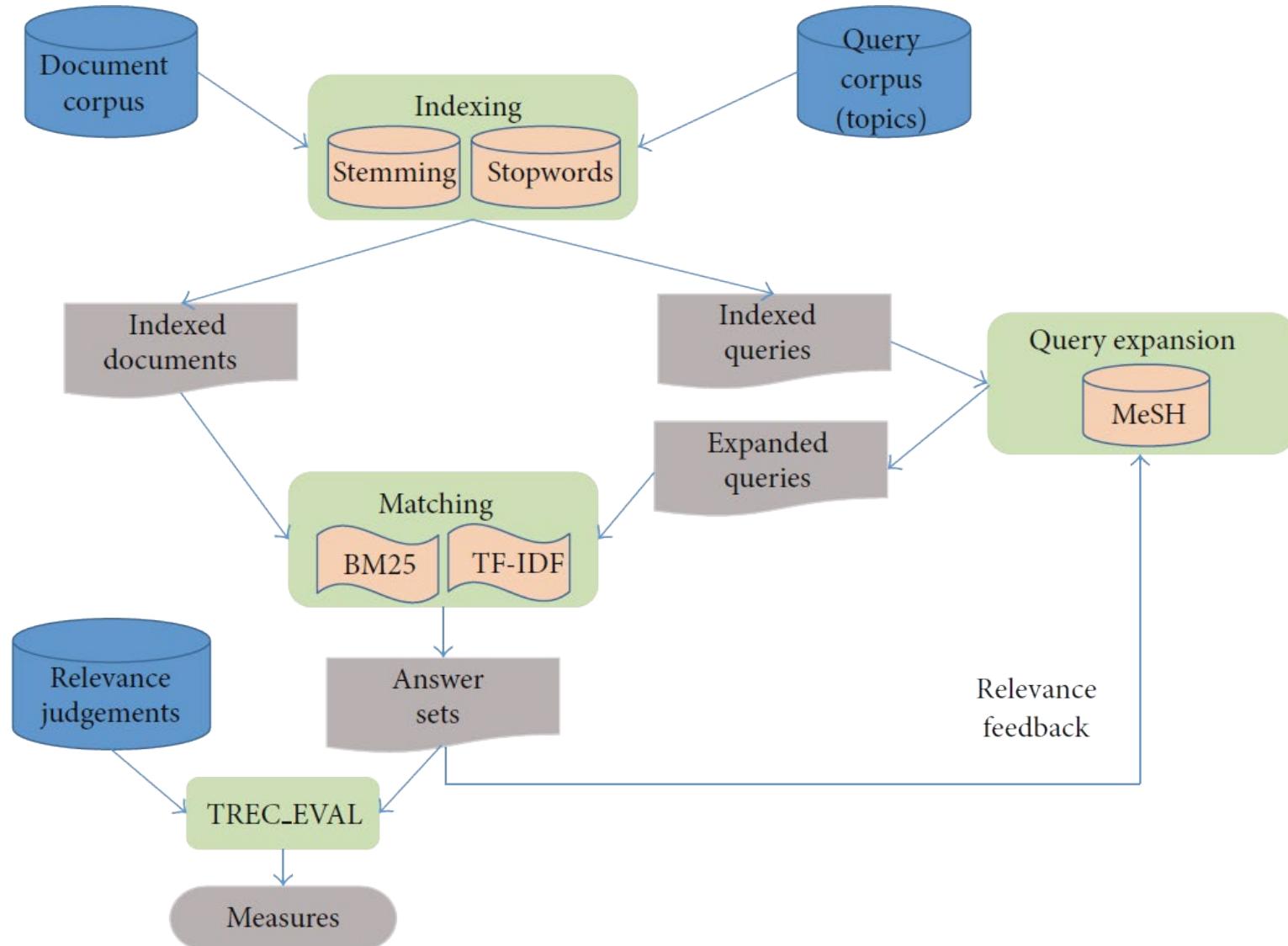
24/03/2016, 11:45

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

TAREFAS USUAIS

1. Análise de sentimento
(classificação)
2. “Named Entity” tagging
3. Tradução
4. “Fill-Mask”
5. Classificação de token
6. Similaridade de sentença
7. Pergunta-resposta – open domain
8. Pergunta-resposta –
“closed domain”
9. Sumarização
10. Classificação “Zero-Shot”
11. Classificação de texto
12. Text2text – geração de
texto
13. Chatbot
14. “Article spinner”



A. R. Rivas, E. L. Iglesias, and L. Borrajo. Study of Query Expansion Techniques and Their Application in the Biomedical Information Retrieval. <http://dx.doi.org/10.1155/2014/132158>

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

TAREFAS PRETENDIDAS

1. Extração das leis e artigos mencionados em uma decisão.
2. Classificação de atos nos diversos tipos processuais (despacho, sentença, acórdão, parecer)
3. Classificação de atos processuais quanto ao desfechos (favorável, desfavorável, multa, resarcimento).
4. Extração da data da sessão, colegiado e relator.
5. Elaboração de ementa.
6. Buscador de precedentes.
7. Buscador de dissídios jurisprudenciais.
8. Elaborador de resenha temática.
9. Buscador de dispositivo legal ou normativo adequado.
10. Identificador de cláusulas restritivas em editais.
11. Identificador de regras de habilitação em editais.

TAREFAS PRETENDIDAS

PERGUNTAS E RESPOSTAS

Domínio – licitações e contratos.

Base 13.000 decisões de mérito do TCE-SP

Tarefa mais próxima QA-Open Domain.

Paradigma – reader-retriever

Vamos usar o bert-large-cased-squad-v1.1-portuguese

Foi treinado a partir do BERTimbau da Neuralmind (<https://neuralmind.ai/>), treinado sobre o SQuAD v1.1 em português do Brasil da Deep Learning Brasil (<http://www.deeplearningbrasil.com.br/>)

TAREFAS PRETENDIDAS

```
!pip install transformers
import transformers
from transformers import pipeline

# fonte: https://pt.wikipedia.org/wiki/Pandemia_de_COVID-19
contexto = r"""

A pandemia de COVID-19, também conhecida como pandemia de coronavírus, é uma pandemia em curso de COVID-19, uma doença respiratória causada pelo coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2). O vírus tem origem zoonótica e o primeiro caso conhecido da doença remonta a dezembro de 2019 em Wuhan, na China. Em 20 de janeiro de 2020, a Organização Mundial da Saúde (OMS) classificou o surto como Emergência de Saúde Pública de Âmbito Internacional e, em 11 de março de 2020, como pandemia. Em 18 de junho de 2021, 177 349 274 casos foram confirmados em 192 países e territórios, com 3 840 181 mortes atribuídas à doença, tornando-se uma das pandemias mais mortais da história. Os sintomas de COVID-19 são altamente variáveis, variando de nenhum a doenças com risco de morte. O vírus se espalha principalmente pelo ar quando as pessoas estão perto umas das outras. Ele deixa uma pessoa infectada quando ela respira, tosse, espirra ou fala e entra em outra pessoa pela boca, nariz ou olhos. Ele também pode se espalhar através de superfícies contaminadas. As pessoas permanecem contagiosas por até duas semanas e podem espalhar o vírus mesmo se forem assintomáticas.

"""

modelo = 'pierreguillou/bert-large-cased-squad-v1.1-portuguese'
nlp = pipeline("question-answering", model=modelo)

pergunta = "Quando começou a pandemia de Covid-19 no mundo?"

resultado = nlp(question=pergunta, context=contexto)

print(f"Resposta: '{resultado['answer']}', score: {round(resultado['score'], 4)}, início: {resultado['start']}, fim: {resultado['end']}")
```

TAREFAS PRETENDIDAS

```
!pip install psycopg
import psycopg

conn = psycopg.connect("dbname=postgres user=postgres password=tceSP#22 " + \
                      "port=5432 " + \
                      "host=ec2-18-231-126-109.sa-east-1.compute.amazonaws.com")

dfResult = pd.read_sql('SELECT TX_PLAIN_TEXT FROM DECISOES_NLP_TC LIMIT 10', conn)

contexto = dfResult['tx_plain_text'][1]

pergunta = "Quando começou a pandemia de Covid-19 no mundo?"

resultado = nlp(question=pergunta, context=contexto)

print(f"Resposta: '{resultado['answer']}' , score: {round(resultado['score'], 4)}, início: {resultado['start']}, fim: {resultado['end']}")
```

TAREFAS PRETENDIDAS

PERGUNTAS E RESPOSTAS

Passo 1 – admitir a pergunta, realizar o embedding, identificar um vetor equivalente.

Passo 2 – identificar no vetor de decisões qual as três mais próximas.

Passo 3 – identificar em cada decisão quais os três parágrafos mais próximos.

Passo 4 – empregar o gerador em cada um dos 9 contextos.

Passo 5 – apresentar a saída e o score ao usuário.

NLP em TC

1. INTRODUÇÃO
2. PRÉ-PROCESSAMENTO
3. REPRESENTAÇÃO DE TEXTO PARA NLP
4. BREVE INTRODUÇÃO A REDES NEURAIS
5. TAREFAS - USUAIS
6. TAREFAS - PRETENDIDAS
7. PIPELINE COMPLETA

TAREFAS PRETENDIDAS

AGENDA DE PASSOS PARA EVOLUIR

Obter dados mais limpos.

Etiquetar dados.

Criar uma tabela de perguntas e respostas.

Treinar word-embeddings para o nosso domínio de linguagem.

Testar.

INTRODUÇÃO

<https://github.com/alexandre-sarquis/NLP>