MCZA017-13
Processamento de Linguagem Natural

Apresentação

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

2Q-2019

Apresentação

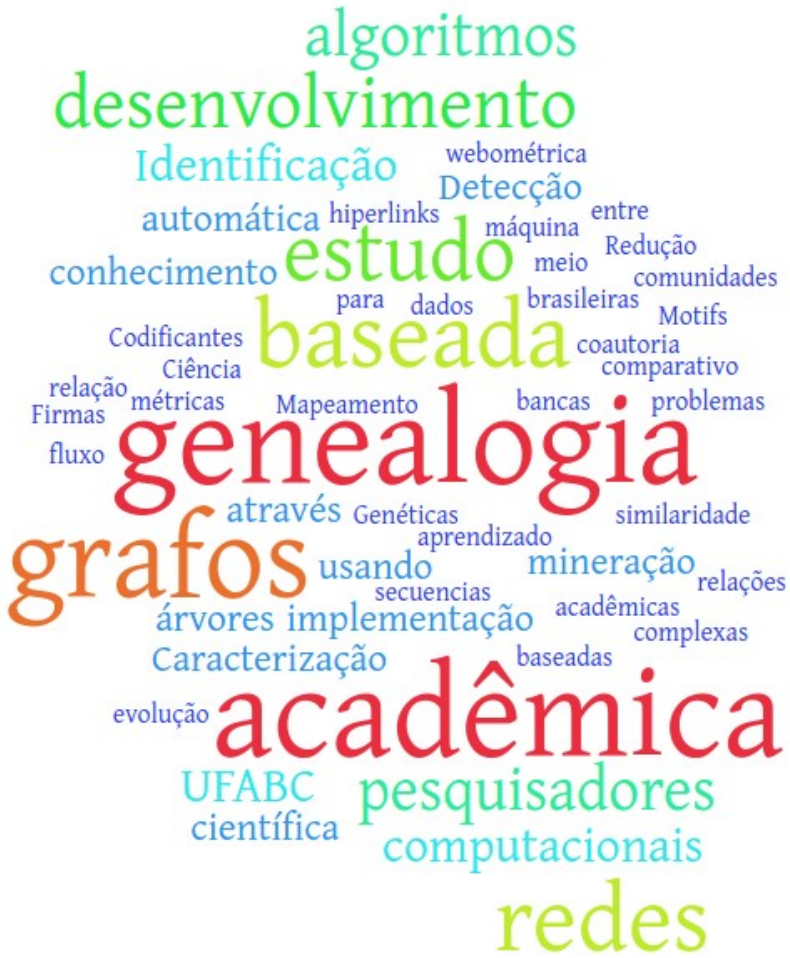
- Jesús P. Mena-Chalco – No CMCC desde 04/2012.

- **Formação:**
  - Engenheiro da Computação.
  - Mestre (2005) e Doutor (2010) em Ciência da Computação.
  Instituto de Matemática e Estatística da USP.

- Sala 517-A, torre 2, 5º Andar.

- **Áreas de pesquisa:**
  - Reconhecimento de padrões, Bibliometria/Cientometria.

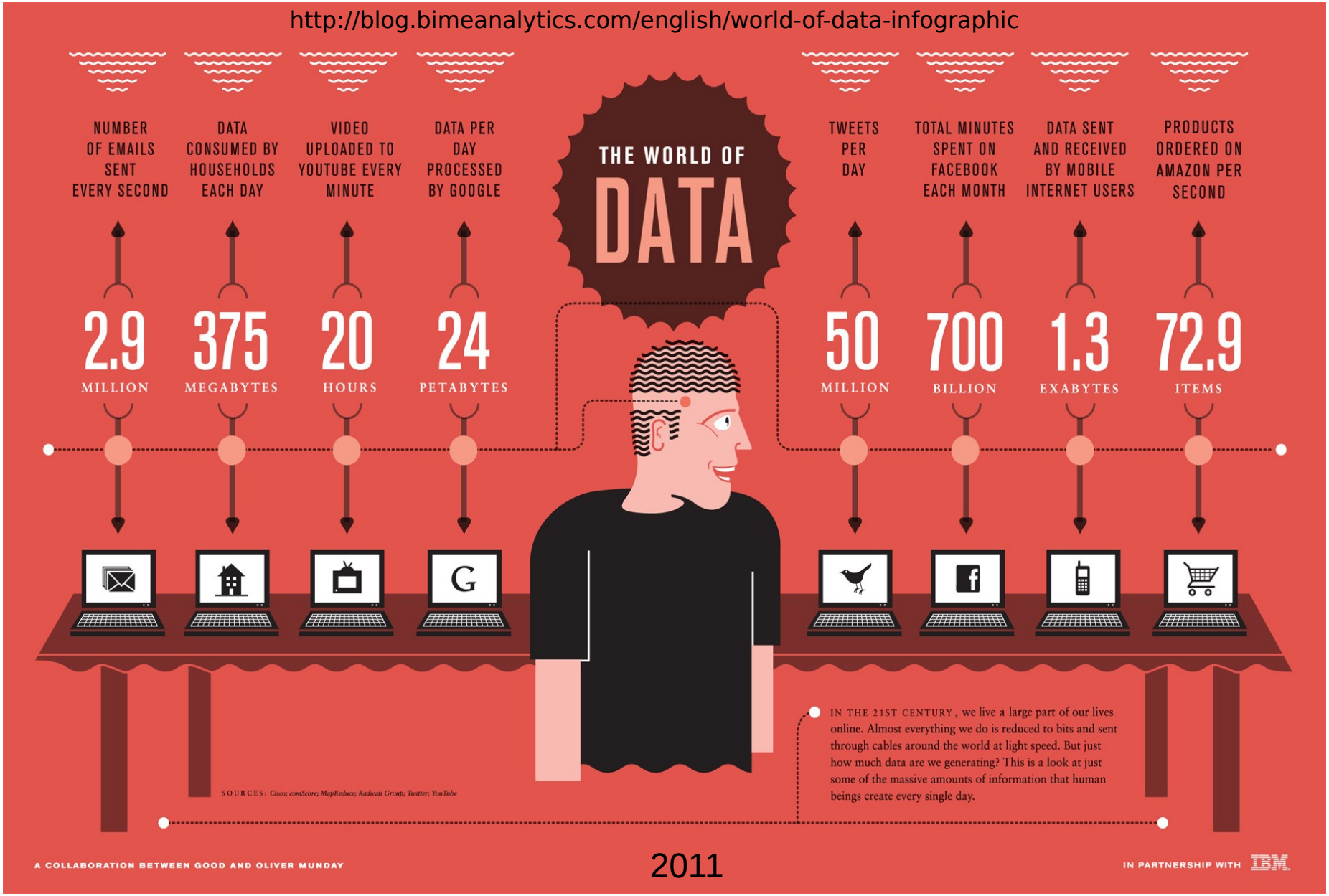
Tópicos de pesquisa



scriptLattes: An open-source knowledge extraction system from the Lattes platform JP Mena-Chalco, RM Cesar-Jr Journal of the Brazilian Computer Society 15 (4), 31-39	192		2009
Identification of protein coding regions using the modified Gabor-wavelet transform J Mena-Chalco, H Carrer, Y Zana, RM Cesar Jr IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 5 …	116		2008
Brazilian bibliometric coauthorship networks JP Mena-Chalco, LA Digiampietri, FM Lopes, RM Cesar Journal of the Association for Information Science and Technology 65 (7 …	78		2014
Minerando e caracterizando dados de currículos lattes L Digiampietri, J Mena-Chalco, J de Jésus Pérez-Alcázar, EF Tuesta, ... Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)	48		2012
A Ciência nas Regiões Brasileiras: Evolução da Produção e das Redes de Colaboração Científica O Sidone, EA Haddad, JP Mena-Chalco Transinformação 28 (1), 1-24	42	*	2016
Scientific communication in Brazil (1998-2012): Indexing, growth, flow and dispersion R Mugnaini, LA Digiampetri, JP Mena-Chalco Transinformação 26 (3), 239-252	30	*	2014
Prospecção de dados acadêmicos de currículos lattes através de scriptLattes JP Mena-Chalco, RMC Junior Bibliometria e Cientometria: reflexões teóricas e interfaces 1, 109-128	30		2013
Brax-ray: an x-ray of the brazilian computer science graduate programs LA Digiampietri, JP Mena-Chalco, POSV de Melo, APR Malheiro, ... Plos One 9 (4), e94541	26		2014
Dinâmica das Relações de Coautoria nos Programas de Pós-Graduação em Computação no Brasil LA Digiampietri, JP Mena-Chalco, GS Silva, LB Oliveira, AP Malheiros, ... Proceedings of BraSNAM	24		2012
Caracterizando as redes de coautoria de currículos Lattes JP Mena-Chalco, LA Digiampietri, RM Cesar-Jr Proceedings of BraSNAM	24		2012
Scholarly publication and collaboration in Brazil: The role of geography OJG Sidone, EA Haddad, JP Mena-Chalco Journal of the Association for Information Science and Technology 68 (1 …	22		2017
Towards automatic discovery of co-authorship networks in the brazilian academic areas JP Mena-Chalco, RMC Junior 2011 IEEE Seventh International Conference on e-Science Workshops, 53-60	21		2011
3D facial expression analysis by using 2D and 3D wavelet transforms SCD Pinto, JP Mena-Chalco, FM Lopes, L Velho, RM Cesar 2011 18th IEEE International Conference on Image Processing, 1281-1284	18		2011

Sobre dados?

http://blog.bimeanalytics.com/english/world-of-data-infographic

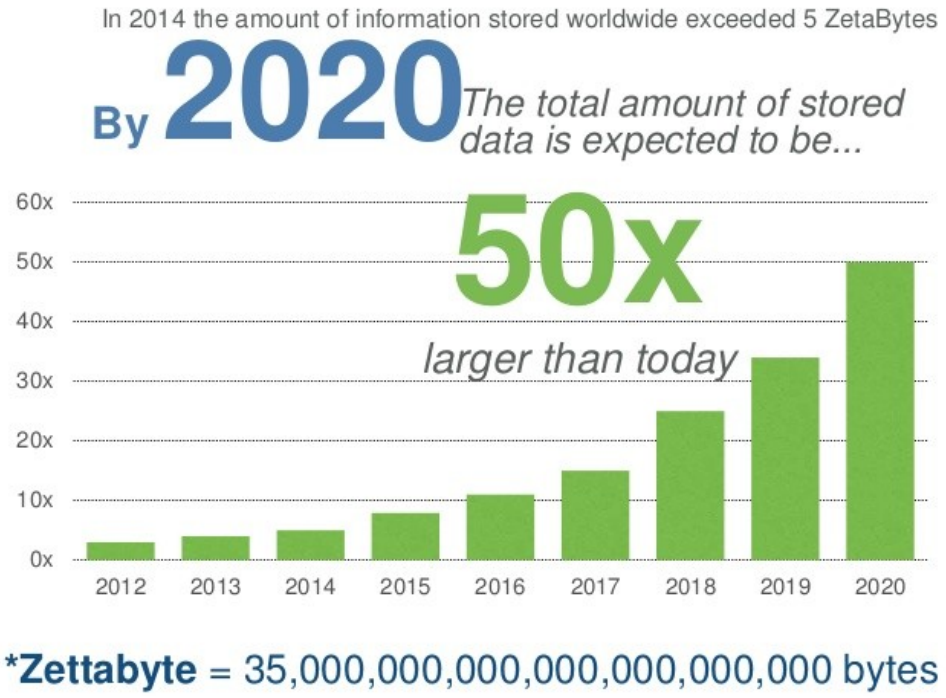


2011

Sobre dados?

source visualcap via @mikequindazzi

A **quantidade** de dados gerados cresce em um ritmo **exponencial**.



(*) Fonte ADEC Group

A grande maioria dos dados nunca será utilizada.

https://www.technologyreview.com/s/530371/big-data-creating-the-power-to-move-heaven-and-earth/

Sobre a disciplina

Ementa

- Introdução ao processamento de linguagem natural.
- Processamento sintático.
- Técnicas de análise (parsing).
- Gramáticas.
- Interpretação semântica.
- Processamento de discurso. Aplicações.

Nas aulas teremos uma introdução sobre os conceitos básicos necessários para a compreensão dos tópicos mais avançados.

Bibliografia

- Jurafsky, D. & Martin, J. (2000). **Speech & language processing**. Pearson Education. 3º edição
- Manning, C. D., & Schütze, H. (1999). **Fundations of statistical natural language processing.** Cambridge: MIT press.
- Koehn, P. (2009). **Statistical machine translation.** Cambridge University Press.
- Steven, B., Klein, E., & Loper, E. (2009). **Natural language processing with python.** OReilly Media Inc.
- Ingersoll, G. S., Morton, T. S., & Farris, A. L. (2013). **Taming text: how to find, organize, and manipulate it.** Manning Publications Co.

Speech and Language Processing (3rd ed. draft)

Dan Jurafsky and James H. Martin

Draft chapters in progress, Sep 23, 2018

This is the release for the start of fall term 2018.

The slides are in the process of being updated now, we are putting them up as we write them.

Significantly rewritten version of 5, 6, 7, 8, 17, 18, 19, 23, 24, 25, and a draft of 9! New pedagogical sequences on neural networks and their training, starting with logistic regression and continuing with embeddings, feed-forward nets, and RNNs. Plus new or improved coverage of BPE, tf-idf, bias in embeddings, beam search decoding, HMMs, connotation frames, lexicon induction. reading comprehension/QA. Some chapters have been moved to the Appendix.

New lecture slides (so far) for chapters 6 and 25.

Here's a single pdf of the whole book-so-far!

Typos and comments welcome (just email slp3edbugs@gmail.com and let us know the date on the draft)!
And feel free to use the draft slides in your classes.

When will the book be finished? We're shooting for late 2019.

(Sorry for the random chapter numbers and occasional missing latex crossrefs in the pdfs, we are constantly reorganizing. The order on this page is our best guess at the final order)

	Chapter	Slides	Relation to 2nd ed.
1:	Introduction		[Ch. 1 in 2nd ed.]
2:	Regular Expressions, Text Normalization, and Edit Distance	Text [pptx] [pdf] Edit Distance [pptx] [pdf]	[Ch. 2 and parts of Ch. 3 in 2nd ed.]
3:	Language Modeling with N-Grams	LM [pptx] [pdf]	[Ch. 4 in 2nd ed.]
4:	Naive Bayes Classification and Sentiment	NB [pptx] [pdf] Sentiment [pptx] [pdf]	[new in this edition]
5:	Logistic Regression		

URLs

- **URL:** http://professor.ufabc.edu.br/~jesus.mena/courses
- **Cadastre-se no Tidia:** http://tidia4.ufabc.edu.br

  Procurar: “PLN-Q2-2019”

Calendário

JUNHO

Dom	Seg	Ter	Qua	Qui	Sex	Sab	
						1	
2	3	4	5	6	7	8	1
9	10	11	12	13	14	15	2
16	17	18	19	20	21	22	3
23	24	25	26	27	28	29	4
30							5

03/06 -Início de 2019.2

20/06 a 22/06 - Corpus Christi

JULHO

Dom	Seg	Ter	Qua	Qui	Sex	Sab	
	1	2	3	4	5	6	5
7	8	9	10	11	12	13	6
14	15	16	17	18	19	20	7
21	22	23	24	25	26	27	8
28	29	30	31				9

08/07 a 09/07 - Revolução Constitucionalista

AGOSTO

Dom	Seg	Ter	Qua	Qui	Sex	Sab
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

20 - Feriado municipal - S B e recesso em S A

- 20 encontros
- 4 aulas dedicadas para apresentações: (22, 26, 28 e 30 de agosto)

Sobre a avaliação

- **(A) Resumos por aula:** → 30%
- **(B) Prova de teoria (única):** 15/08 → 40%
- **(C) Projeto (relatórios+apresentação):** → 30%
- Prova substitutiva: 30/08
- Prova de recuperação: Q3/2019

Obs: Para aprovar na disciplina não pode reprovar em nenhum dos 3 quesitos (A,B,C).

Atribuição de conceitos:

A: nota ≥ 9

B: 7,5 ≤ nota < 9

C: 6 ≤ nota < 7,5

D: 5,0 ≤ nota < 6

F: nota < 5,0

Sobre a avaliação

- **Resumos por aula:**
  - Redação de 250 a 500 palavras (apenas texto sem formato).
  - Envio pelo Tidia (prazo máx. 48h após cada aula).
  - Todos os resumos serão publicados na pág. da disciplina.

- **Prova de teoria (única):** 15/08
  - Serão abordados os conceitos vistos em aula.

- **Projeto (relatórios+apresentação):**
  - Mini-relatório 1 (1 página – 10%): 27/06
  - Mini-relatório 2 (3 páginas – 20%): 25/07
  - Mini-relatório 3 (5 páginas – 50%): 19/08
  - Apresentações orais (15min – 20%): 22, 26, 28 e 30/08

Sobre a projeto

Estudo e implementação de um artigo científico relacionado com um ou mais tópicos de Processamento de Linguagem Natural.

Restrições sobre a escolha do artigo:

- Publicado após 2010.
- Número de citações maior ou igual a 6 (indicar fonte).
- Número de páginas maior ou igual a 6.
- O código fonte não deve estar disponível.

Observações:

- Grupos de 4 pessoas (descrever o trabalho de cada membro).
- Use a linguagem Python na implementação.
- O código fonte e dados devem ser disponibilizados.

Sobre a linguagem de programação

- **Usaremos Python**
  - Em aulas vamos ter a parte prática nas quintas-feiras.

- Instale o **Jypiter Notebook** no seu computador
  http://jupyter.org/install

- Usaremos, preferencialmente, os conceitos básicos da linguagem:
  - Listas
  - Matrizes
  - Dicionários