

CURSO: Ciência de Dados e Inteligência Artificial – 2º semestre de 2023
DISCIPLINA: **Processamento de Linguagem Natural e Sistemas de Recuperação de Informação**
PROFESSOR(ES): Alexandre Rademaker
CARGA HORÁRIA: 60h
PRÉ-REQUISITO: Estrutura de Dados e Algoritmos
HORÁRIO E SALA DE ATENDIMENTO: segundas-feiras 12-14h
SALA: sala professor na EMap

PLANO DE ENSINO

1. Ementa

Este curso apresenta as técnicas fundamentais de processamento de linguagem natural e de sistemas de recuperação de informação. Tem como objetivo explicar o potencial e as principais limitações destas técnicas. Alguns problemas de pesquisa atuais são introduzidos e algumas aplicações discutidas e avaliadas. Os alunos também serão apresentados à tarefas prática em processamento de linguagem natural.

Introdução à PLN; morfologia de estados finitos; part-of-speech tagging; gramáticas; estruturas de dependências; semântica composicional e semântica lexical; semântica distribucional;; pesquisas em PLN.

Recuperação de Informação; construção de índices; avaliação da recuperação: precisão e revocação, coleções de referência; consultas: consultas lógicas, consultas ordenadas por relevância, estruturas de acesso ao vocabulário, busca sequencial no vocabulário, busca exata e aproximada.

2. Objetivos da disciplina

Ao final do curso, os alunos devem: 1) ser capaz de discutir aplicações e limitações no processamento de linguagem natural; 2) conhecer as técnicas fundamentais de processamento de linguagem e tarefas como processamento morfológico, análise sintática e semântica de textos; 3) compreender como essas técnicas se baseiam e se relacionam com outras áreas da matemática e as áreas de pesquisa existentes.

3. Procedimentos de ensino (metodologia)

trabalhos em grupos e individuais e avaliações individuais.

4. Conteúdo programático detalhado

Datas	Tópico	Atividades
07-09/08	Introdução PLN e IR	

14-16/08	Demo QA, bases de conhecimento, RDF/Web Semântica	
21-23/08	RDF/SPARQL, DBPedia, WikiData, Triple Stores, IR	
28-30/08	IR (índices invertidos, queries booleanas), Apache Solr	
04-06/09	Expressão regular, distância de edição, segmentação	
11-13/09	Modelos de linguagem, N-grams, classificação	
18-20/09	Morfologia, Foma	
25-27/09	POS tagging, NER (rotulação de sequências)	
	A1	
11/10	Sintaxe, Gramáticas, PCFG	
16-18/10	Dependências Universais	
23-25/10	Semântica Composicional	
30-01/11	Análise Semântica: AMR, MRS, Propbank	
06-08/11	Semântica Lexical	
13/11	Semântica Distribucional	
22/11	Tópicos Extra	
	A2	

5. Procedimentos de avaliação

Trabalhos semanais em grupo ou individuais irão compor A1 e A2.

$$A1 = Pr * 0.4 + MA * 0.5 + Pa * 0.1$$

Onde Pr é a prova escrita, MA é a média dos trabalhos passados até a respectiva avaliação e Pa é nota de participação.

6. Bibliografia Obrigatória

- Jurafsky, D. & Martin, J. (2008) *Speech and language processing*. Prentice Hall. <https://web.stanford.edu/~jurafsky/slp3/>
- Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.
- Steven Bird, Ewan Klein and Edward Loper. *Natural Language Processing with Python*. O'Reilly. 2009 <https://www.nltk.org/book/>

7. Bibliografia Complementar

-
- Processamento de Linguagem Natural, Conceitos, Técnicas e Aplicações em Português. <https://brasileiraspln.com/livro-pln/1a-edicao/>
 - Croft, W. B., Metzler, D., & Strohman, T. (2009). Search Engines: Information Retrieval in Practice (1 ed.). Addison Wesley.
 - Ingwersen, P., & Järvelin, K. (2005). The Turn: Integration of Information Seeking and Retrieval in Context. Springer.
 - Nitin Indurkha and Fred J. Damerau. Handbook of Natural Language Processing, Second Edition. Chapman & Hall/CRC. 2010
 - Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Addison Wesley.
 - Alexander Clark and Chris Fox. The Handbook of Computational Linguistics and Natural Language Processing. Wiley 2012
 - Natural Language Understanding, James Allen
 - <https://www.amazon.com/Natural-Language-Understanding-James-Allen/dp/0805303340>
 - Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax
<https://www.morganclaypool.com/doi/abs/10.2200/S00493ED1V01Y201303HLT020>
 - Linguistic Fundamentals for Natural Language Processing II: 100 Essentials from Semantics and Pragmatics
<https://www.morganclaypool.com/doi/abs/10.2200/S00935ED1V02Y201907HLT04>
 - 3**
 - Finite State Morphology
<https://press.uchicago.edu/ucp/books/book/distributed/F/bo3613750.html>
 - Natural Language Processing in Lisp: An Introduction to Computational Linguistics

-
- <https://www.amazon.com/Natural-Language-Processing-Lisp-Computational/dp/0201178257>
 - <https://staff.fnwi.uva.nl/d.j.n.vaneijck2/cs/>
 - https://olzama.github.io/CL-seminar/annotated_bib.html
 - Nitin Indurkha and Fred J. Damerau. Handbook of Natural Language Processing, Second Edition. Chapman & Hall/CRC. 2010
 - Matthew Honnibal and Patrick Harrison. Deep Learning with Text: A Modern Approach to Natural Language Processing with Python and Keras. O'Reilly, 2018

8. Minicurrículo do(s) Professor(s)

Alexandre também é professor adjunto na Escola de Matemática Aplicada da Fundação Getúlio Vargas (EMAp / FGV) e Pesquisador na IBM, no passado, foi também professor na PUC-Rio. Graduou-se na UFRJ em Computação, fez o mestrado na UFF com especialização de métodos formais e sistemas distribuídos e doutorado na PUC em computação, Lógica Matemática. Alexandre foi coordenador da Comissão Especial de Processamento de Linguagem da SBC (CE-PLN), e é membro da comissão da Global Wordnet Association. Com mais de 100 publicações em eventos e periódicos internacionais, atua em projetos de processamento de linguagem, construção de recursos linguísticos (corpus com Universal Dependencies para PT, Wordnet do Português etc). As principais áreas de pesquisa de Alexandre são: lógica, representação de conhecimento, processamento de linguagem natural e programação funcional.

9. Link para o Currículo Lattes

<http://lattes.cnpq.br/0675365413696898>