

PROCESSAMENTO DE LINGUAGEM NATURAL (IMD1107)

Carga Horária: 04 Créditos (60 horas)

Semestre: 2024.2

Docente: Elias Jacob de Menezes Neto

Horários: 3N1234 (19:00 às 22:10, sala A303)

Quantidade de vagas: 30

1. Objetivo

- Introduzir os conceitos básicos e aplicações de Processamento de Linguagem Natural (PLN)
- Apresentar técnicas para tarefas de aprendizado de máquina que utilizam PLN de modo supervisionado e não supervisionado
- Propiciar experiência prática com ferramentas e *frameworks* de PLN.

2. Ementa

Overview sobre os desafios de PLN. Corpos linguísticos anotados para Inglês e Português, aprendizado supervisionado e avaliação da eficácia. Módulos, componentes, tarefas de um sistema de PLN: pré-processamento; processamento ao nível das palavras; processamento ao nível das estruturas sintáticas ou de dependências; processamento ao nível semântico, pragmático, de discurso, resolução de coreferência. Representações distribuídas para palavras (vector space models, word embeddings). Técnicas de Aprendizado de Máquina para PLN. Técnicas de aprendizado profundo para PLN. Aplicações de PLN diversas.

3. Metodologia

Oferecerei o curso no formato de ensino presencial. A rotina do curso é a seguinte: (1) a primeira atividade de cada aula consiste na realização da chamada, (2) seguida da revisão do conteúdo da aula anterior; (3) a etapa seguinte focará no conteúdo do dia.

Utilizarei um método de ensino *top-down*, que é diferente de como a maioria dos cursos funciona. Normalmente, em uma abordagem *bottom-up*, você primeiro aprende todos os componentes separados que irá utilizar e, em seguida, reúne-os gradualmente em estruturas mais complexas. Os problemas com isso é que os alunos frequentemente perdem a motivação, não têm uma visão do "quadro geral" e não sabem o que precisarão lá na frente. Sabe aquela sensação de "pra que danado eu vou usar isso?", pois então.

David Perkins, professor de educação da Universidade de Harvard, tem um livro chamado "[Making Learning Whole](#)", em que ele usa o jogo de beisebol como analogia (pense em futebol, dá no mesmo). Quando crianças estão aprendendo a jogar, não exigimos que elas memorizem todas as regras do jogo e entendam todos os detalhes técnicos antes de permitir que elas joguem. Ao contrário, elas começam a jogar com apenas uma noção geral e, gradualmente, aprendem mais regras/detalhes à medida que o tempo passa.

Tudo isso para dizer, não se preocupe se você não entender tudo no início. É esperado que você não entenda. Começaremos a usar algumas "caixas pretas" que ainda não foram explicadas, e então, iremos nos aprofundar nos detalhes delas mais tarde. **Concentre-se no que as coisas FAZEM, não no que elas SÃO.**

Você não vai aprender nada só assistindo a aula. As pessoas aprendem **fazendo** (construindo código) e **explicando** o que aprenderam (apresentando e ajudando os outros).

Na linha do David Perkins, utilizaremos cadernos Jupyter com código em python durante praticamente todas as aulas. Os cadernos são feitos de modo que você consiga estudar tudo depois com calma. Tudo fica mais fácil quando vira código.

Cadernos Jupyter utilizados: <https://jacob.al/IMD1107>

Fora da sala de aula, estarei à disposição para tirar dúvidas mediante agendamento via e-mail elias.jacob@ufrn.br. Caso você tenha críticas, elogios ou comentários que deseje expressar de forma anônima, você pode usar o seguinte formulário do Google Forms <https://jacob.al/feedbacks>

4. Procedimentos de avaliação da aprendizagem

Considero a avaliação da aprendizagem como um processo dinâmico que abarca aquisição de conteúdos, habilidades e desenvolvimento de atitudes. Durante o curso, os alunos(as) serão avaliados individualmente e em grupo. As avaliações individuais serão fruto da participação em encontros e prova escrita. No caso de avaliação em grupo, serão realizadas duas atividades para discussão de conteúdos adicionais e implementação de soluções. A ideia é trabalhar pontos além daqueles vistos no semestre, com uma ideia de *Bring Your Own Data*, isto é, com datasets diferentes daqueles utilizados em sala de aula.

A nota ao final do curso será a média aritmética das notas obtidas durante cada unidade (10,0). As atividades avaliativas de unidades encontram-se descritas no quadro da agenda do curso.

5. Cronograma e critérios para a realização das atividades e validação da assiduidade dos discentes

Apresento o cronograma com os critérios para a realização das atividades ao final deste documento. A assiduidade dos estudantes é avaliada pela presença nos encontros síncronos. Para ser aprovado, o estudante deve obter 75% de presença e média parcial igual ou superior a 7,0 (sete) ou média parcial igual ou superior a 6,0 (seis), com rendimento acadêmico igual ou superior a 4,0 (quatro) em todas as unidades.

6. FAQ

- Eu vou precisar faltar em razão de {doença do cachorro, doença própria, guerra nuclear etc.}. Preciso avisar? **Não**
- Se eu chegar depois da chamada, vou levar falta? **Sim (casos excepcionais serão tratados excepcionalmente)**
- Quantas vezes eu posso faltar? **Quantas o SIGAA disser que pode.**
- Posso ter a minha falta abonada? **Não existe abono de faltas.**
- As provas têm consulta? **Não.**

7. Agenda do curso

Semana	Data	Aula	Atividades pontuadas
UNIDADE 1 <i>Back to basics</i> Noções básicas de Aprendizado de Máquina e Processamento de Linguagem Natural			
1	17/09	Apresentação da disciplina e revisão geral de aprendizado de máquina	-
24/09 NÃO HAVERÁ AULA			
2	01/10	Introdução ao Processamento de Linguagem Natural. Principais tarefas de PLN. Representando textos com vetores.	-
3	08/10	<i>Corpora</i> para tarefas de PLN em português e inglês. Ferramentas de anotação.	-
4	15/10	Módulos, componentes e tarefas de um sistema de PLN.	-
5	22/10	Atividade avaliativa 1 – <i>Bring your own data</i> – Apresentação de projeto com PLN (grupos de 3 pessoas – envio no SIGAA)	10,00
UNIDADE 2 <i>Let there be words (and machines)</i> Processamento de Linguagem Natural e Aprendizado de Máquina Aplicados			
6	29/10	Modelagem de tópicos.	-
7	05/11	Classificação de texto com aprendizado de máquina. Análise de sentimentos.	-
8	12/11	Classificação de tokens com aprendizado de máquina	-
9	19/11	Atividade avaliativa 2 – <i>Bring your own data</i> – Apresentação de projeto com PLN (grupos de 3 pessoas – envio no SIGAA)	10,00
Unidade 3 <i>How deep the rabbit hole goes?</i> Processamento de Linguagem Natural e Aprendizado Profundo			
10	26/11	<i>Transfer learning</i> e o <i>big bang</i> na área de PLN	
11	03/12	Similaridade sintática e semântica com aprendizado profundo	
12	10/12	Redes recorrentes	
13	17/12	Transformers	
14	07/01	Agentes inteligentes e pipelines envolvendo grandes modelos de linguagem	
		Prova escrita individual – nota 3ª unidade	
15	14/02	Reposição	10,00
Nota final: média das três unidades			
16	19/06	Reposição	10,00

8. Referências básicas

BENGFORT, B.; BILBRO, R.; OJEDA, T. **Applied Text Analysis with Python: Enabling Language-Aware Data Products with Machine Learning**. [s.l.] O'Reilly Media, 2018.

BROWNLEE, J. **Deep Learning for Natural Language Processing: Develop Deep Learning Models for your Natural Language Problems**. [s.l.] Machine Learning Mastery, 2017.

HOWARD, J.; GUGGER, S. **Deep Learning for Coders with fastai and PyTorch**. [s.l.] O'Reilly Media, 2020.

LANE, H.; HAPKE, H.; HOWARD, C. **Natural Language Processing in Action: Understanding, analyzing, and generating text with Python**. [s.l.] Manning Publications, 2019.

OLIVE, J.; CHRISTIANSON, C.; MCCARY, J. (EDS.). **Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation**. New York, NY: Springer New York, 2011.

9. Referências complementares

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. [s.l.] MIT Press, 2016.

RAO, D.; MCMAHAN, B. **Natural Language Processing with PyTorch: Build Intelligent Language Applications Using Deep Learning**. [s.l.] O'Reilly Media, 2019.

SRINIVASA-DESIKAN, B. **Natural language processing and computational linguistics: a practical guide to text analysis with Python, Gensim, spaCy, and Keras**. Birmingham Mumbai: Packt, 2018.

TRASK, A. W. **Grokking deep learning**. Shelter Island: Manning, 2019.

TUNSTALL, L.; VON WERRA, L.; WOLF, T. **Natural Language Processing with Transformers: Building Language Applications with Hugging Face**. [s.l.] O'Reilly Media, Incorporated, 2022.

VAJJALA, S. et al. **Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems**. [s.l.] O'Reilly Media, 2020.