Universidade de São Paulo
Faculdade de Filosofia, Letras e Ciências Humanas
Departamento de Ciência Política

FLS 6513
Processamento de Língua Natural Aplicada para Ciência Política e Análise de Políticas Públicas (*Natural Language Processing for Political and Policy Analysis: A Machine Learning Approach*)

&

FLP0478
Processamento de Língua Natural (NLP) Aplicada para Ciência Política e Análise de Políticas Públicas

Lorena G. Barberia

Coordenador de Laboratórios:
Pedro Schmalz

2º semestre / 2024
Turma Vespertino: 14:00-18:00
Turma Noturno: 19:00-23:00

As redes sociais e a *World Wide Web* fornecem uma rica fonte de dados detalhados que refletem a evolução dos posicionamentos políticos ao longo do tempo e em resposta a vários eventos. Este curso baseado em projetos enfoca a interseção entre aprendizado de máquina, processamento de língua natural e ciência política, mostrando como as estruturas de Aprendizado de Máquina (*Machine Learning*) e PLN (Processamento da Língua Natural) podem ser aplicadas para responder a questões importantes envolvendo política e políticas públicas. Um componente-chave do curso é um projeto de pesquisa de um semestre usando um grande conjunto de dados de texto. Os alunos irão formular, explorar e abordar exercícios focados usando esses dados. Ao final do curso, os alunos terão adquirido experiência prática no uso de dados em larga escala e soluções eficazes de PNL.

Objetivos

- Aprender as habilidades básicas de programação em Python e como aplicá-las no Processamento de Língua Natural (PLN).
- Compreender os conceitos fundamentais de aprendizado profundo e como LLMs tais como o BERT, (Bidirectional Encoder Representations from Transformers), LLaMA, podem ser utlilzados para PLN.
- Explorar diferentes métodos de classificação em PLN, como classificação binária, multiclasse e multirrótulo. Além disso, como são criados os bancos para a classificação de textos, e alguns cuidados metodológicos que devem ser tomados.
- Aplicar esses conceitos e técnicas em projetos práticos para classificação de textos, incluindo a criação de modelos de classificação de posicionamento e sentimento.

Avaliação

1. Entrega das listas semanais de Exercícios. 60% da nota final.
   Os alunos deverão entregar listas semanais de exercícios que trabalham os conteúdos discutidos nos laboratórios em sala e nos textos obrigatórios.

2. Trabalho Final (Pós-graduação somente). 40% da nota final.
   O aluno deverá escolher um problema de classificação de interesse e tentar aplicar as técnicas aprendidas no curso utilizando o banco de dados utilizado durante a disciplina.

Programação, Objetivos e Referências:

1. Introdução

Grimmer J, Roberts ME, Stewart BM (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Capítulos 1 & 2

Lab 1 : Introdução: Apresentação do curso e introdução ao básico de Python

Objetivos:

- Introduzir o Google Colab, principal ferramenta de nossas aulas
- Apresentar os principais tipos de variáveis e estruturas de dados em Python

Referências para o Lab:

Menezes, Nilo Ney Coutinho (2019). Capítulo 3 - "Variáveis e entrada de dados", Capítulo 4 - "Condições", Cap. 5 - "Repetições" e Capítulo 6 - "Listas" em "Introdução à Programação com Python". Terceira Edição, Editora Novatec. pp. 49-136.

McKinney, Wes (2022) - Chapter 3. "Built-In Data Structures, Functions, and Files" in "Python for Data Analysis Data Wrangling with pandas, NumPy, and Jupyter". Third Edition. O'Reilly.

Matthes, Eric (2023) - Python Crash Course - A Hands-On, Project-Based Introduction to Programming. 3rd Edition. No Starch Press.

2. Seleção e Representação: A Construção do Corpus

Grimmer J, Roberts ME, Stewart BM (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Capítulos 3 & 4

Lab 2: Condicionais e Loops

Objetivos:

- Apresentar algumas declarações condicionais
- Introduzir as estruturas de repetição (Loops) e suas duas principais formas

Referências para o Lab:

Matthes, Eric (2023) - "Chapter 5: If Statements" and "Chapter 7: User Input and While Loops" in "Python Crash Course - A Hands-On, Project-Based Introduction to Programming". 3rd Edition. No Starch Press.

Lutz, Mark (2013). Chapter 12 "if Tests and Syntax Rules", Chapter 13 "While and for Loops", and Chapter 14 "Iterations and Comprehensions" in "Learning Python". 5th Edition, O"Reilly Media, Inc.

3. Mensuração e Aprendizado Supervisionado

Grimmer J, Roberts ME, Stewart BM (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Capítulos 15, 17 e 18

Lab 3: Funções, Módulos e Introdução ao Numpy

Objetivos:

- Apresentar as funções e como criá-las
- Como importar módulos, convenções de importação e principais módulos
- Introdução ao Numpy e seus principais módulos

Referências para o Lab:

McKinney, Wes (2022) - Chapter 3. "Built-In Data Structures, Functions, and Files" and Chapter 4. "NumPy Basics: Arrays and Vectorized Computation" in "Python for Data Analysis Data Wrangling with pandas, NumPy, and Jupyter". Third Edition. O'Reilly.

Lutz, Mark (2013). Chapter 16 "Function Basics", Chapter 22 "Modules: The Big Picture", Chapter 23 "Module Coding Basics" in "Learning Python". 5th Edition, O"Reilly Media, Inc.

Matthes, Eric (2023) - "Chapter 8: Functions" in "Python Crash Course - A Hands-On, Project-Based Introduction to Programming". 3rd Edition. No Starch Press.

4. Conceitos Básicos de Estatística para Aprendizado de Máquina Supervisionado I

Cerulli, Giovanni. (2023) *Fundamentals of Supervised Machine Learning: With Applications in Python, R, and Stata*. Springer Nature. Capítulo 3

Lab 4: Pandas, Dados e Estatísticas Resumidas

Objetivos:

- Introduzir o aluno ao Pandas
- Como abrir arquivos Excel, .CSV, etc.
- Manipulação, Organização e Limpeza de Dados
- Estatísticas Resumidas

Referências:

McKinney, Wes (2022) - Chapter 5. "Getting Started with pandas", Chapter 6. "Data Loading, Storage, and File Formats", Chapter 7. "Data Cleaning and Preparation" and "8. Data Wrangling: Join, Combine, and Reshape" in "Python for Data Analysis Data Wrangling with pandas, NumPy, and Jupyter". Third Edition. O'Reilly.

5. Conceitos Básicos de Estatística para Aprendizado de Máquina Supervisionado II

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer. Capítulo 2. Statistical Learning.

Lab 5. Visualização de Dados em Python

Objetivos:

- Introduzir o aluno aos principais pacotes de visualização no Python
- Identificação das melhores visualizações para cada tipo de dado
- Destacar algumas visualizações específicas de Text as Data

Referências:

McKinney, Wes (2022) - Chapter 9. "Plotting and Visualization", Chapter 10. "10. Data Aggregation and Group Operations" in "Python for Data Analysis Data Wrangling with pandas, NumPy, and Jupyter". Third Edition. O'Reilly.

Massaron, Lucas; Boschetti, Alberto (2019) - Chapter 5 "Visualization, Insights, and Results" in "Python Data Science Essentials - A practitioner's guide covering essential data science principles, tools, and techniques." Third Edition. Published by Packt Publishing Ltd.

6. Classificação

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer. Capítulo 4. Classification.

Grimmer J, Roberts ME, Stewart BM (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press. Capítulos 19 e 20

Lab 6. O problema de Classificação e suas métricas

Objetivos:

- Compreender as principais abordagens para a classificação de dados
- Introdução de métodos de reamostragem

Referências:

Géron, Aurelien. (2022). "Classification" in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems"

Kuhn, Max and Johnsonn, Kjell. (2018). "Chapter 11 Measuring Performance in Classification Models" in "Applied Predictive Modeling". Springer Science+Business Media, New York.

7. Resampling

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer. Capitulo 5: Resampling.

Lab 7: Overfitting, Reamostragem e Validação dos Resultados

Objetivos:

- Alertar para evitar a confiança total na precisão como métrica dos seus resultados
- Apresentar métodos que garantam a estabilidade dos resultados
- Cross-Validation e Test Sets

Referências:

Kuhn, Max and Johnsonn, Kjell. (2018). "Chapter 4 Over-Fitting and Model Tuning" in "Applied Predictive Modeling". Springer Science+Business Media, New York.

8. Support Vector Machines and Decision Trees

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). "8 Tree-Based Methods" and "9 Support Vector Machines" in *An Introduction to Statistical Learning: with Applications in Python*. Springer.

Lab 8: Métodos de Classificação I : SVM e Decision Trees

Objetivos:

- Conhecer os principais métodos de classificação utilizados no aprendizado de máquina
- Analisar diversos desfechos em saúde a partir do uso de diferentes métricas usadas em epidemiologia, saúde global e ciências políticas
- Identificar limitações metodológicas para análise de políticas de saúde

Referências:

Géron, Aurelien. (2022). "5. Support Vector Machines" e "6. Decision Trees" in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems"

Kuhn, Max and Johnsonn, Kjell. (2018). "Chapter 13 Nonlinear Classification Models " in "Applied Predictive Modeling". Springer Science+Business Media, New York.

9. Random Forests

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). "8 Tree-Based Methods" in *An Introduction to Statistical Learning: with Applications in Python*. Springer.

Lab 9: Métodos de Classificação II: Mais árvores e Ensemble

Objetivos:

- Apresentar outros modelos Tree-Based
- Método Ensemble

Referências:

Géron, Aurelien. (2022). "7. Ensemble Learning and Random Forests" in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems"

Kuhn, Max and Johnsonn, Kjell. (2018). "Chapter 14 Classification Trees and Rule-Based Models " in "Applied Predictive Modeling". Springer Science+Business Media, New York.

10. Introdução ao Deep Learning

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). "10 Deep Learning" in *An Introduction to Statistical Learning: with Applications in Python*. Springer.

Lab 10: Introdução ao Deep Learning

Objetivos:

- Introduzir o que é o aprendizado profundo e como ele funciona
- Conceitos principais de aprendizagem profunda como learning rate, epochs, gradient descent, etc.

Referências:

Géron, Aurelien. (2022). "10. Introduction to Artificial Neural Networks with Keras" and "11. Training Deep Neural Networks" in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems".

11. Introdução a BERT

Lab 11. Classificação de Textos com BERT I

Objetivos:

- Apresentar o BERT e como ele opera na classificação de textos

- Destacar principais cuidados a se tomar na criação de um banco de dados

Referências:

Géron, Aurelien. (2022). "2. End-to-End Machine Learning Project" in "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems".

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2 [cs.CL].

12. BERT II

Lab 12: Classificação de Textos com BERT II - Pré-Processamento, Data-loader e Separação dos Dados

Objetivos:

- O que é o pré-processamento e como fazê-lo para o BERT
- Como criar o Data-Loader (utilizando Classes)
- Separando os bancos de treino e de teste, definindo a Cross-Validation

Referências:

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2 [cs.CL]..

13. BERT III

Lab 13: Classificação de Textos com BERT III - Construindo o Loop de Treinamento e Hiperparâmetros

Objetivos:

- Como construir o Loop de Treinamento e Validação do BERT
- Extraindo métricas de treinamento e validação
- Definir os hiperparâmetros de treinamento

Referências:

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2 [cs.CL].

14. BERT IV

Lab 14. Classificação de Textos com BERT IV - Treinando e Apresentando seus resultados

Objetivos:

- Quais são os seus resultados e como se comparam com os de outras pesquisas

- Montando a visualização e apresentação dos seus resultados

15. Novas Ferramentas: LLMaAA

Zhang, R. et al. (2023) 'LLMaAA: Making Large Language Models as Active Annotators', arXiv [cs.CL]. Available at: http://arxiv.org/abs/2310.19596.

16. Novas Ferramentas: Sabiá, Cabrita e Bode

Pires, R. et al. (2023) 'Sabiá: Portuguese Large Language Models', in Intelligent Systems. Springer Nature Switzerland, pp. 226–240. Available at: https://doi.org/10.1007/978-3-031-45392-2_15.

Material Complementar:

Tutorial de Introdução ao Python - PET Tele (2009).

Documentação do Python - Disponível em https://docs.python.org/3/library/stdtypes.html

Projeto Panda do IME-USP. Disponível em https://panda.ime.usp.br/panda/default/intro

Curso Gratuito da DSA sobre Python. Disponível em https://www.datascienceacademy.com.br/cursosgratuitos

Referências

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing Perspective. Peer J. Comput. Sci. 10.

Araújo, P.H.L., Campos, T.E., Braz, F.A.; Silva, N.C.: Victor. (2020). A Dataset for Brazilian Legal Documents Classification. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 11–16 May, pp. 1449–1458. Marseille.

Barberá, P., Rivero, G. (2014). Understanding the Political Representativeness of Twitter Users. *Social Science Computer Review* (Vol. 33, Issue 6, pp. 712–729). Sage Publications.

Barberia, Lorena Guadalupe; Schmalz, Pedro Henrique de Santana; Roman, Norton Trevisan. When Tweets Get Viral - A Deep Learning Approach for Stance Analysis of Covid-19 Vaccines Tweets by Brazilian Political Elites. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14. , 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023 . p. 104-114. DOI: https://doi.org/10.5753/stil.2023.233961.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media, Inc.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805v2 [cs.CL].

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer.

Géron, Aurelien. (2022). “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.”

Grimmer J, Roberts ME, Stewart BM (2022). *Text as Data: A New Framework for Machine Learning and the Social Sciences*. Princeton University Press.

Jurafsky, D. & Martin, J. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd Edition draft.

Kaufman, Aaron Russell, Peter Kraft, e Maya Sen. (2019). "Improving Supreme Court Forecasting using Boosted Decision Trees." *Political Analysis* 27 (3): 381–87.

Katz, D.M., Bommarito, M.J.I., Blackman, J. (2014). Predicting the behavior of the Supreme Court of the United States: A general approach. In: arXiv e-prints, page arXiv:1407.6333 (2014).

Kuhn, Max and Johnsonn, Kjell. (2018). “Applied Predictive Modeling”. Springer Science+Business Media, New York.

Liu, Bing. (2010). Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*
Lutz, Mark (2013). “Learning Python”. 5th Edition, O”Reilly Media, Inc.

Massaron, Lucas; Boschetti, Alberto (2019) - “Python Data Science Essentials - A practitioner's guide covering essential data science principles, tools, and techniques.” Third Edition. Packt Publishing Ltd.

Matthes, Eric (2023) - Python Crash Course - A Hands-On, Project-Based Introduction to Programming. 3rd Edition. No Starch Press.

Zhang, R. et al. (2023) ‘LLMaAA: Making Large Language Models as Active Annotators’, arXiv [cs.CL]. Available at: http://arxiv.org/abs/2310.19596.