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
//
FLP 0442 – Política Comparada V

2º semestre / 2023

Aulas: Segundas-feiras 14:00-18:00 e 19:30-23:00

Professora Lorena G. Barberia

As redes sociais e a *world wide web* fornecem uma rica fonte de dados detalhados que refletem a evolução dos sentimentos políticos ao longo do tempo e em resposta a vários eventos de notícias. Este curso será ministrado no formato *project-based* buscando introduzir os alunos a interseção entre aprendizado de máquina, processamento da língua natural e ciência política, mostrando como as estruturas de AM e PLN podem ser aplicadas para responder a questões importantes envolvendo a política e as políticas públicas. O curso está organizado em exercícios aplicados baseados em um projeto de pesquisa usando um grande conjunto de dados de texto relevantes que foram coletados sobre uma questão de grande relevância no contexto da pandemia da COVID-19 e o papel das elites políticas na formulação de políticas públicas. Os alunos irão formular, explorar e abordar exercícios focados nestes dados. Ao final do curso, os alunos terão adquirido experiência prática no uso de dados em larga escala e soluções eficazes de PLN para examinar questões políticas.

Pré-requisitos: Alunos de pós-graduação devem ter completado os cursos de Métodos I (FLS 5028) e Métodos Quantitativos II (FLS 6183) ou equivalentes. Alunos de graduação devem ter completado cursos de inferência estatística (FLP 0406 ou equivalente) e econometria.

Monitores: Guilherme Tiengo, Rebeca Carvalho e Pedro Schmalz

Visão Geral do Curso

1. Visão geral do curso e introdução ao Python (5 semanas)
2. Introdução ao aprendizado estatístico e aprendizado de máquina (1 semana)
3. Problema de Classificação e Métodos de Reamostragem (2 semanas)

4. Visão geral dos métodos de classificação (3 semanas)
5. BERT para Classificação de Texto (4 semanas)

Avaliação

1. Entrega das Listas Semanais (100% para a Graduação / 60% para a Pós) - Os alunos serão avaliados com base na entrega e correção das listas semanais.

2. Trabalho Final (Pós-Graduação Somente - 40% da Nota Final) - Alunos de pós-graduação farão um trabalho final tentando aplicar as técnicas aprendidas ao longo do curso em um banco de dados de seu interesse.

Material Complementar:

Tutorial de Introdução ao Python - PET Tele (2009). Disponível em: https://www.telecom.uff.br/pet/petws/downloads/tutoriais/python/tut_python_2k100127.pdf

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

Programa e Objetivos

Aula 1 – 07 de agosto de 2023

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

Leitura Obrigatória:

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.

Bibliografia Complementar:

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.

Aula 2 – 14 de agosto de 2023

Condicionais e Loops

Objetivos:

- Apresentar algumas declarações condicionais

- Introduzir as estruturas de repetição (Loops) e suas duas principais formas.

Leitura Obrigatória:

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.

Bibliografia Complementar:

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.

Aula 3 – 21 de agosto de 2023

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.

Leitura Obrigatória:

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.

Bibliografia Complementar:

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.

Aula 4 – 28 de agosto de 2023

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

Leitura Obrigatória:

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.

Aula 5 – ***04 de setembro de 2023***

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

Leitura Obrigatória:

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.

Bibliografia Complementar:

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.

Aula 6 - 11 de setembro de 2023

Aprendizado Estatístico e Aprendizado de Máquina

Objetivos:

- Apresentar o aprendizado de máquina e principais tarefas;

- Diferenciar Aprendizado Supervisionado de Não Supervisionado;

- Possíveis problemas que podem surgir e cuidados a se tomar

Leitura Obrigatória:

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2022). "Introduction" e "2 Statistical Learning" in *An Introduction to Statistical Learning: with Applications in R*. Springer.

Bibliografia Complementar:

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

Aula 7- 18 de setembro de 2023

O problema de Classificação e suas métricas

Objetivos:

- Compreender as principais abordagens para a classificação de dados

- Introdução de métodos de reamostragem

Leitura Obrigatória:

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2022). "Classification" in *An Introduction to Statistical Learning: with Applications in R*. Springer.

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

Bibliografia Complementar:

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

Aula 8 - 25 de setembro de 2023

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

Leitura Obrigatória:

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2022). " Resampling Methods" in *An Introduction to Statistical Learning: with Applications in R*. Springer.

Bibliografia Complementar:

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

Aula 9 - 02 de outubro de 2023

Métodos de Classificação I : SVM e Decision Trees

Objetivos:

- Conhecer os principais métodos de classificação utilizados no aprendizado de máquina

Leitura Obrigatória:

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*

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2022). "8 Tree-Based Methods" and "9 Support Vector Machines" in *An Introduction to Statistical Learning: with Applications in R*. Springer.

Bibliografia Complementar:

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

Aula 10 - 09 de outubro de 2023

Métodos de Classificação II: Mais árvores e Ensemble

Objetivos:

- Apresentar outros modelos Tree-Based;

- Método Ensemble

Leitura Obrigatória:

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*

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2022). “8 Tree-Based Methods” in *An Introduction to Statistical Learning: with Applications in R*. Springer.

Bibliografia Complementar:

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

Aula 11 - *16 de outubro de 2023*

Métodos de Classificação III: 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.

Leitura Obrigatória:

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*.

Gareth James, Gareth; Witten, Daniela; Hastie, Trevor and Tibshirani, Robert. (2022). “10 Deep Learning” in *An Introduction to Statistical Learning: with Applications in R*. Springer.

Aula 12 - *23 de outubro de 2023*

Classificação de Textos com BERT I - Construção do Banco de Dados;

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;

- Stance Vs. Sentiment

Leitura Obrigatória:

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].

Barberia, Lorena; Schmalz, Pedro; Roman, Norton (2023). “When Tweets Get Viral - A Deep Learning Approach for the Sentiment Analysis of Covid-19 Vaccines Tweets of Brazilian Political Elites.” *Working Paper.*

Barberia, Lorena; Rosa, Isabel Seelaender Costa; Carvalho, Rebeca de Jesus; Schmalz, Pedro Henrique de Santana; Paula, Gustavo Fernandes de; Garibe, André Colucci André Colucci (2023). *Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020. Codebook*.

Aula 13 - *30 de outubro de 2023*

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.

Leitura Obrigatória:

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].

Barberia, Lorena; Schmalz, Pedro; Roman, Norton (2023). “When Tweets Get Viral - A Deep Learning Approach for the Sentiment Analysis of Covid-19 Vaccines Tweets of Brazilian Political Elites.” *Working Paper.*

Barberia, Lorena; Rosa, Isabel Seelaender Costa; Carvalho, Rebeca de Jesus; Schmalz, Pedro Henrique de Santana; Paula, Gustavo Fernandes de; Garibe, André Colucci André Colucci (2023). *Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020. Codebook*.

Aula 14 - ***06 de novembro de 2023***

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.

Leitura Obrigatória:

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].

Barberia, Lorena; Schmalz, Pedro; Roman, Norton (2023). “When Tweets Get Viral - A Deep Learning Approach for the Sentiment Analysis of Covid-19 Vaccines Tweets of Brazilian Political Elites.” *Working Paper.*

Barberia, Lorena; Rosa, Isabel Seelaender Costa; Carvalho, Rebeca de Jesus; Schmalz, Pedro Henrique de Santana; Paula, Gustavo Fernandes de; Garibe, André Colucci André Colucci (2023). *Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020. Codebook*.

Aula 15 - ***13 de novembro de 2023***

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.

Leitura Obrigatória:

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].

Barberia, Lorena; Schmalz, Pedro; Roman, Norton (2023). “When Tweets Get Viral - A Deep Learning Approach for the Sentiment Analysis of Covid-19 Vaccines Tweets of Brazilian Political Elites.” *Working Paper.*

Barberia, Lorena; Rosa, Isabel Seelaender Costa; Carvalho, Rebeca de Jesus; Schmalz, Pedro Henrique de Santana; Paula, Gustavo Fernandes de; Garibe, André Colucci André Colucci (2023). *Mapping Political Elites COVID-19 Vaccine Tweets in Brazil in 2020. Codebook*.