OFICINA 2: Processamento de Linguagem Natural aplicada à Gestão Pública

Professor: Samuel de Souza Barbosa

Trilha: Dados, Monitoramento e Avaliação

Carga-horária: 10 horas/aula

Breve currículo: Mestre em Economia pela EPGE/FGV-RJ, e mestrando em Estatística pela UFMG (2021-2022). Graduado em Administração Pública pela Fundação João Pinheiro (2010). Especialista em Políticas Públicas e Gestão Governamental desde 2011, tendo atuado em áreas como Monitoramento e Avaliação, Orçamento e Finanças Públicas e Processamento e Análise de Dados. Experiência e interesse em ciência de dados, modelagem matemática/estatística e aprendizado de máquinas.

Currículo Lattes: http://lattes.cnpq.br/7423848539417847

Servidor(a) Público(a) do Executivo Estadual: Sim

Período de realização: 30 de maio; 01, 06, 08 e 15 de junho

Modalidade (se presencial ou híbrido): Presencial

Pré-requisitos exigidos para matrícula: Noções básicas de lógica de programação.

Ementa: Processamento de Linguagem Natural: conceitos, ferramentas e aplicações. Tratamento e análise de documentos e de dados textuais. Reconhecimento de Entidades Nomeadas. Similaridade textual e Análise de Semântica Latente. Modelagem de tópicos e Alocação Latente de Dirichlet.

Objetivo: Familiarizar os alunos com técnicas estatísticas para análise textual. Documentos e textos estão presentes em todas as áreas e atividades da Administração Pública. Ao adotar métodos computacionais tratar textos como dados, é possível aplicar análises estatísticas como o reconhecimento de menções a pessoas, lugares e organizações, de tópicos em documentos e a identificação de documentos similares.

Método Didático: Aulas expositivas e aplicações computacionais.

Programa:

1. **Aula 1: Introdução ao Processamento de Linguagem Natural**

   Introdução aos principais conceitos e aplicações do Processamento de Linguagem Natural (NLP – *Natural Language Processing*).

2. **Aula 2: Ferramentas de processamento de texto**

   Uso da linguagem Python para tratamento de dados textuais. Pacotes, métodos e funções utilizadas no processamento de texto. Expressões regulares. Distância entre palavras ou expressões.

3. **Aula 3: Texto enquanto dados: métodos para análise**

Métodos para a análise de textos sob a ótica estatística e computacional. Frequência de termos e de documentos, matrizes termo-documento e termo-termo. Matriz TF-IDF. Vetores de palavras.

4. **Aula 4: Similaridade Textual**

   Cálculo de similaridade entre documentos. Análise de Semântica Latente.

5. **Aula 5: Reconhecimento de entidades nomeadas e Modelagem de Tópicos**

   Reconhecimento de menção a pessoas, lugares e organizações. Identificação de tópicos. Alocação Latente de Dirichlet.

Critérios para obtenção de certificado:

Frequência em pelo menos 4 encontros.

Cronograma de aulas proposto:

Aula	Dia	Horário	Modo
Introdução ao Processamento de Linguagem Natural	30/05	18:00	Presencial
Ferramentas de processamento de texto	01/06	18:00	Presencial
Texto enquanto dados: métodos para análise	06/06	18:00	Presencial
Similaridade Textual	08/06	18:00	Presencial
Reconhecimento de entidades nomeadas e Modelagem de Tópicos	15/06	18:00	Presencial

Referências básicas:

Jurafsky, D., & Martin, J. H. (2000). Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Referências complementares:

Pedregosa *et al.* (2011). Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.