

## **Análise de redes e vínculos utilizando base textuais**

**(90 min)**

**Descrição e Objetivos:** Processamento de Linguagem Natural (NLP, do inglês Natural Language Processing) é uma área de pesquisa e aplicações dedicada a explorar como computadores podem ser usados para compreender e manipular a linguagem natural em textos e discursos e realizar tarefas com isso. Pesquisadores de NLP visam reunir o conhecimento de como seres humanos entendem e usam a linguagem de maneira que ferramentas e técnicas adequadas possam ser desenvolvidas.

Por sua vez, a Ciência de Redes (NS, do inglês Network Science) é o campo do conhecimento que estuda redes complexas, presentes em diferentes aspectos da vida cotidiana: telecomunicações, biologia, cognitiva, semântica, sociologia, considerando elementos distintos e as conexões existentes entre eles.

As duas áreas ganharam bastante destaque nos últimos anos com o advento das redes sociais e o aumento exponencial de material produzido através delas, além do avanço da inteligência artificial e o poder computacional disponível. Unir as técnicas estudadas pelos dois campos oferece possibilidades para compreensão situacional antes, se não impossíveis, bastante dificultadas. Por exemplo, entender os relacionamentos entre pessoas e organizações apenas lendo artigos jornalísticos é uma tarefa árdua para humanos devido ao grande número de informações e possibilidades. Ao unir o técnicas de reconhecimento de entidades nomeadas provindas do NLP com a construção de redes e a utilização de medidas de centralidade, graus de importância, detecção de comunidades... Todo o contexto pode ser traçado, compreendido e exposto.

O objetivo desse workshop/tutorial é mostrar a partir de ferramentas *open-source/cross-platform* de NLP e NS como utilizar uma metodologia para analisar vínculos a partir de bases textuais, com especial foco em textos jornalísticos, mas com técnicas que podem ser estendidas para outras bases, sendo assim focado para uma audiência ampla.

**Metodologia:** o Workshop/Tutorial consiste em duas etapas:

A. Parte conceitual (Apresentação com Slides – 40 min): i) Consolidação de conceitos de

NLP; Estado da arte de NLP, principais ferramentas ii) Consolidação de conceitos de NS; Principais Algoritmos, Principais Medidas, Identificação de Comunidades

B. Parte Prática (50 min): Nesta parte será mostrado passo-a-passo as etapas para realizar uma análise. Os participantes poderão acompanhar com os próprios computadores ou utilizar o material como futura referência. As etapas mostradas serão:

- a. Como instalar o ambiente de trabalho
  - i. Anaconda + Jupyter
  - ii. Gephi
  - iii. Módulos de Python: spaCy e Networkx
- b. Carregar base de dados textual e executar algoritmo de reconhecimento de entidades;
- c. Criar um grafo (rede complexa) utilizando rotinas em Python;
- d. Gerar uma visualização para verificação do resultado;
- e. Conexão com o Gephi e utilização para execução de algoritmos de maneira visual.

**Público-alvo:** Interessados em tecnologia, noções de programação básica recomendada, mas não é requisito. Voltado para profissionais de comunicação/jornalistas e profissionais com formação em ciências exatas, como cientistas de computação, matemáticos e engenheiros, sem experiência em Processamento de Linguagem Natural e/ou Ciência de Redes.

É recomendável ao participante trazer seu próprio laptop.

**Instrutor:** Fernando Guimarães Ferreira - cientista de dados e sócio-fundador da TWIST, formado em Engenharia Eletrônica e de Computação pela UFRJ e doutor em Inteligência Computacional pelo Programa de Engenharia Elétrica da COPPE. Participou da colaboração CERN-UFRJ por 8 anos, trabalhando no desenvolvimento de diversos sistemas, estando alocado no laboratório durante quatro vezes. A TWIST encontra-se no Parque Tecnológico da UFRJ e é especializada em integrar dados não-estruturados e aliá-los à inteligência computacional. Lattes: <http://lattes.cnpq.br/6518221600690415>

**Email:** [Fernando.ferreira@twistsystems.com](mailto:Fernando.ferreira@twistsystems.com)

**Websites de referências:**

- <https://anaconda.org/>

- <https://gephi.org/>
- <http://jupyter.org/>
- <https://spacy.io/>
- <http://www.twistsystems.com/pt-br/>

## Referências:

- Al Omran, Fouad Nasser A., and Christoph Treude. "Choosing an NLP library for analyzing software documentation: a systematic literature review and a series of experiments." *Proceedings of the 14th International Conference on Mining Software Repositories*. IEEE Press, 2017.
- Barabási, Albert-László. *Network science*. Cambridge university press, 2016.
- Bastian, Mathieu, Sebastien Heymann, and Mathieu Jacomy. "Gephi: an open source software for exploring and manipulating networks." *Icwsn* 8 (2009): 361-362.
- Benikova, Darina, et al. "Network of the day: Aggregating and visualizing entity networks from online sources." *Proc. NLP4CMC Workshop at KONVENS, Hildesheim, Germany*. 2014.
- Bonzanini, Marco. *Mastering social media mining with Python*. Packt Publishing Ltd, 2016.
- Chowdhury, Gobinda G. "Natural language processing." *Annual review of information science and technology* 37.1 (2003): 51-89.
- Kochtchi, Artjom, T. von Landesberger, and Chris Biemann. "Networks of Names: Visual Exploration and Semi-Automatic Tagging of Social Networks from Newspaper Articles." *Computer Graphics Forum*. Vol. 33. No. 3. 2014.
- Ragan-Kelley, M., et al. "The Jupyter/IPython architecture: a unified view of computational research, from interactive exploration to communication and publication." *AGU Fall Meeting Abstracts*. 2014.
- Sarkar, Dipanjan. *Text Analytics with Python: A Practical Real-World Approach to Gaining Actionable Insights from Your Data*. Apress, 2016.