KDD 2016 Tutorials

Big Natural Language Data Processing

Gabor Melli / OpenGov, Inc.

Matt Seal / OpenGov, Inc.

Tutorials

The automated processing of large volumes of text data has become a mission critical capability in a wide-range of industries. Current tools enabled data scientists to produce every more impactful NLP systems. Getting these systems started however, can still be challenging.

This tutorial will review the best-practice data-driven methods, tools, and resources for many common applications that require the processing of high volume, high velocity and/or mixed veracity textual data. We will show PDF content
extraction, distributed ETL patterns, and NLP tooling with working code samples (mostly in Python) using the Spark and AWS Lambda distributed computing platforms.

When completed participants will understand and be able to prototype components of an end-to-end NLP system that achieve baseline results. They will also be shown advanced strategies that can be expanded and further explored
without re-implementing the underlying baseline infrastructure.We assume that audience members have a general understanding of textual data, NLP applications, and data science methods.