Session Title: Analyzing Textual Data

Instructor(s) Name: Steven Bedrick

Instructor(s) Information:

Steven Bedrick has a PhD biomedical informatics and is an assistant professor at the Center for Spoken Language Understanding at OHSU.

Audience: Librarians and researchers new to data science

Audience level: Beginner

Session Day/Time: Wednesday, November 08, 2017, 10am-12pm.

Session Description:
Textual data is everywhere, and all too often is thought of as an analytical "black hole." Surveys with free-text responses (or the dreaded "Other - please specify"), news articles, interview transcripts, books, clinical records: all are potentially useful sources of information, but can be challenging to work with. In this session, we will discuss the basics of how to process unstructured and semi-structured textual data.

Participants will learn about different types of analyses that can be easily conducted on textual data, as well as basic techniques for matching patterns and characterizing corpora. Participants will also learn about basic issues of character encoding and Unicode.

Session Learning Objectives
A participant will, at the end of the session, be able to:

- Understand the basic steps of a text-mining analysis (tokenization, etc.)
- Use regular expressions to:
    - Match patterns in text
    - Extract information from matches
- Use simple text processing techniques to explore and compare texts
- Explain why "résumé" sometimes appears as "rÃ©sumÃ©" in emails

Course Materials and Supplies
Laptop & Web Browser

R/RStudio (if you want to follow along)

Required
- Choose a favorite book from Project Gutenberg (link below)
- If you plan on following along, install the `tidytext` R package ahead of time

Resources
- Julia Sigle and David Robinson's "Text Mining with R"
- Steven Bird, Ewan Klein, and Edward Loper's "Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit"
- Project Gutenberg