DATA WORKFLOWS AND NETWORK ANALYSIS

Professor Christopher Church
University of Nevada, Reno
christopherchurch@unr.edu

Overview

This workshop will discuss methods of data retrieval, data cleaning, and visualization. Participants will discuss how websites are structured and learn how to collect a data set with webscraping. Participants will learn how to use tools like OpenRefine for cleaning and transforming data and then visualize data using Gephi, an open source tool for network analysis.

Software Required

- Google Chrome
- Microsoft Excel (or OpenOffice Calc)
- Gephi 0.8.2-beta
- OpenRefine 2.5
- Texteditor (preferably Notepad++)
- Latest Java Installation 8u51

Suggested Readings

- Borgatti, S. P., A. Mehra, D. J. Brass, and G. Labianca. 2009. Network analysis in the social sciences. *Science* 323:892-895.
- "Tooling up for the Digital Humanities: Data Visualization." Stanford University.
- Verborgh, Ruben and Max De Wilde "Chapter 2." *Using OpenRefine, The Book*.
- Wainer, Howard. "Introduction, Chapter 1, Chapter 2." *Graphic Discovery.* Princeton, 2007.
- Wetherell, Charles. 1998. "Historical Social Network Analysis". *International Review of Social History.* 43 (6): 125.
- Weingart, Scott. "Demystifying Networks." *Journal of Digital Humanities,* Vol. 1, No. 1, 2011.

Professor Christopher Church
University of Nevada, Reno
christopherchurch@unr.edu

Schedule

Monday

9-11

Introductions (30 minutes)
- *Introduction to class (10 minutes)*
- *Participant Introductions (20 minutes)*

Data Visualization: Potential and Limits (30 minutes)
- *PowerPoint (30 minutes)*

Introduction to Networks (60 minutes)
- *PowerPoint on Networks (40 minutes)*
- *Making Your First Network List, Matrix and Graph by Hand (20 minutes)*

1-4

Installing Gephi and Learning the Interface (75 minutes)
- *Downloading and Installing Gephi (15 minutes)*
- *Overview of the Interface (30 minutes)*
- *Overview of Layouts (30 minutes)*

Building a Network of Class Participants (45 minutes)
- *Making the CSV file (10 minutes)*
- *Importing the Nodes/Edges (10 minutes)*
- *Layouts, Sizing, Colors (20 minutes)*
- *Exporting (10 minutes)*
- *Discussing the Limits (10 minutes)*

Project Work (60 minutes)

Tuesday

9-11

OpenRefine Overview (80 minutes)
- *Installing OpenRefine (30 minutes)*
- *Basics Overview – Facets, Clustering, Duplicates, Transformations (20 minutes)*
- *REGEX and GREL PowerPoint and Activity (30 minutes)*

Sample Data Cleaning Activity – Powerhouse Museum (40 minutes)
- *Importing and Examining (10 minutes)*
- *Cleaning (30 minutes)*

1-4

The Participant Network Returns (120 minutes)
- *Importing Participant Network (15 minutes)*
- *Cleaning our Participant Network (30 minutes)*
- *Exporting the Participant Network (15 minutes)*
- *Visualizing with Gephi – Attributes, New Networks, and New Information (60 minutes)*

Professor Christopher Church
University of Nevada, Reno
christopherchurch@unr.edu

Project Work (60 minutes)

Wednesday

2-4

Analyzing Networks Basics (60 minutes)

PowerPoint on Structure, Centrality, and Network Statistics (30 minutes)
Activity with Facebook Network (30 minutes)

Project Time (60 minutes)

Thursday

9-11

Analyzing Networks: Structural Analysis (60 minutes)

PowerPoint: Components/Isolates, Small Worlds, Preferential Attachment (30 minutes)
Activity with Citation Network (25 minutes)

Analyzing Networks Intermediate: Community Detection and Clustering (60 minutes)

PowerPoint (25 minutes)
Activity with Forest Gump Network (10 minutes)
Activity with Classroom Network (30 minutes)

1-4

Where to get Network Data? Scraping vs. API Consumption (30 minutes)

Structure of the Web Overview PowerPoint – XML/JSON vs Edge Lists (30 minutes)

APIs, REST, and Getting Some Data (90 minutes)

Set up Twitter API and use Hurl.It to get Twitter Data (45 minutes)
Using OpenRefine and Excel to Extract Tweet Hashtags and Mentions (45 minutes)

Project Time (60 minutes)

Friday

9-11

Visualize Twitter Network in Gephi (120 minutes)

Formatting Consumed Twitter Data (10 minutes)
Importing into Gephi (10 minutes)
Analyzing the Network (30 minutes)
Choosing a Layout (20 minutes)
Drawing Some Conclusions (30 minutes)
Exporting your Graph (20 minutes)

1-4

Giving your graphs meaning with INKscape (100 minutes)

PowerPoint – Importance of Scales and Vector Files (15 minutes)
Downloading and Installing INKscape (25 minutes)
Overview of INKscape (15 minutes)
Importing Graph File to INKscape (10 minutes)
Deciding Upon / Creating a Scale (30 minutes)
Saving/Exporting (5 minutes)

Professor Christopher Church
University of Nevada, Reno
christopherchurch@unr.edu

Conclusion/Summary of Week and Q&A (30 minutes)
Project Time (50 minutes)