William J Turkel

History 9877A: Digital Research Methods (Fall 2013)

Historical research now crucially involves the acquisition and use of digital sources. In History 9877A, students learn to find, harvest, manage, excerpt, cluster and analyze digital materials throughout the research process, from initial exploratory forays through the production of an electronic article or monograph which is ready to submit for publication.

Course Description

Three Scenarios

Why should you take a course in digital research methods?

1. You’ve just returned from a whirlwind trip to the archives. On your laptop you have about nine thousand digital photographs of various documents. You could spend the next few years going

through the pictures one at a time and typing notes into a word processor. Or you could write a small script to convert each image into readable text and drop the whole batch into a custom search engine. In less than an hour you could be searching for words and phrases anywhere in your primary sources.

2. You discover that the Internet Archive has a collection of eight hundred online texts that are directly related to your research. You could look through the list of titles in your web browser and click on the links one at a time, scanning each to see if it is relevant. Even if you cut-and-paste notes from the sources to a word processor, it will still take you at least a few months to go through the collection. Or you could write a small script to download all of the sources to your own machine and run a clustering program on them. This sorts the texts into folders of closely related documents, then subdivides those by topic. In less than an hour, you would be able to visualize the contents of the whole collection and focus in on the topics that are of immediate interest to you.

3. You've been working with the written corpus of a historically significant figure. You have the books and essays that he or she wrote, their diary entries and their correspondence with a large number of other individuals. How do you make sense of a lifetime of writing? Can you chart important changes in someone's conceptual world? Spot the emergence of new ideas in the discourse of a community? Map the ever-changing social relations between a network of correspondents?

In this course you will learn to apply techniques that are currently used by fewer than one percent of working historians. Computation won't magically do your research for you, but it will make you *much* more efficient. You can focus on close reading, interpretation and writing, and use machines to help you find, summarize, organize and visualize sources.

Prerequisites, Workload, Blogging and Evaluation

There are no prerequisites for the course other than a willingness to learn new things and the perseverance to keep working when you're confused or when you realize that you could spend a lifetime learning about the topics and technologies that we will cover in class, and still not master them all. Students will come into the course with very different levels of experience and expertise. Some, probably most, will be familiar only with the rudiments of computer and internet use. A few may already be skilled programmers.

This course also requires that you spend at least a little bit of time each day (say 20-30 minutes) practicing your new skills. It's a lot like learning a new language, learning to play a musical instrument or going to the gym. It is going to be hard at first, but be patient with yourself and ask a lot of questions. With daily practice, you will soon find ways to do your research and coursework faster and more efficiently. If you can't commit to regular practice, however, you should probably not take this course. The techniques that you learn in this class build cumulatively week-by-week. In addition to regular practice, it is essential that you attend every meeting of the class and do the readings carefully.

Every student in the class will have an academic blog and will be required to make weekly posts to it. These entries do not have to be long (300-500 words per week is ample). The use of blogging is to encourage you to engage in 'reflective practice,' that is, to force you to think about your learning and research as you are doing it. It also provides me with feedback for how the course is going. You can use each week's blog entry to talk about what you learned, things that were clear or not, things you would like to know how to do, and so on.

Before the first class you should go to either WordPress or Blogger (not both) and create an account and a blog. If possible, create the blog under your own name; if not, choose something professional sounding. Post an introductory message about yourself and then send me the URL of your blog so that I can add you to the course blogroll for History 9877A.
You will be graded on your participation in class (20%) and on your reflective blogging (80%). There will be no midterm or final examinations, and no final paper.

Text and Requirements
There is one required text for this course.

Shotts, William E., Jr. *The Linux Command Line: A Complete Introduction*. No Starch Press, 2012.
In addition, you will need a computer that you can use daily (ideally a laptop that you can bring to every class) and a USB flash drive (16 Gb or larger).

GitHub Repository
https://github.com/williamjturkel/Digital-Research-Methods

Week 01: Basic Text Analysis

- Overview
  - We can get some idea of what a text is about by studying the frequency of words in it
- Readings
  - Shotts Ch 1, What is the Shell?
  - Shotts Ch 2, Navigation

    - Shotts Ch 3, Exploring the System
- In-class activity
    - Starting and stopping a VirtualBox virtual machine
    - Basic Text Analysis with Command Line Tools in Linux
- In-class discussion
    - Taxonomy of digital sources
        - text (ASCII, Unicode)
        - markup (XML, HTML, TEI)
        - human-readable and machine-readable
        - open vs. closed / proprietary formats
        - documents (MS Word, PDF)
        - … and lots more we will discuss as course continues
    - Operating systems
        - Windows and Mac (built on Unix)
        - Linux (descended from Unix)
        - Dual-boot machines
        - Virtual machines
- Linux commands to study and practice
    - **clear**, **date** and **cal**
    - **wget**
    - **ls**
    - **pwd** and **cd**
    - **head**, **tail** and **less**
    - **file**
    - **wc**
    - **tr**
    - redirection operators: **<**, **>**, **|**

Week 02: Pattern Matching and Permuted Term Indexing

- Recap and Overview
    - Word frequencies can give us some idea of what a text is about
    - We can search for particular words or expressions in a text using *regular expressions*, a powerful pattern matching language
    - We can see how a particular word is used by building a permuted index (also called a *concordance* or a *keyword in context* (KWIC) listing)
- Readings

  - Shotts Ch 4, Manipulating Files and Directories
  - Shotts Ch 5, Working with Commands
  - Shotts Ch 6, Redirection
- In-class activity
  - Create folder for week01 and clean up previous week's work
  - Pattern Matching and Permuted Term Indexing with Command Line Tools in Linux
- In-class discussion
  - rtfm
  - Googling the error
  - Useful sites: Linux Documentation Project, Debian Books,Stack Overflow
- Linux commands to study and practice
  - **mkdir**
  - **cp**, **mv** and **rm**
  - filename wildcards
  - character classes (try using these with **tr** and **egrep** too)
  - **type**, **which**, **man**, **apropos**, **whatis**
  - **wget**
  - **cat**, **head**, **tail** and **less**
  - **sort**
  - **uniq**
  - **ptx**
  - redirection operators: **<**, **>**, **>>**, **|**
  - */dev/null*

Week 03: Batch Downloading and Building Simple Search Engines

- Recap and Overview
  - We can download any online text and begin to make sense of it by analyzing word frequencies, searching for regular expressions and building a concordance
  - We can download arbitrarily large collections of sources automatically
  - Unlike our previous methods, a *search engine* can return sources ranked by *relevance* to a particular query
- Readings
  - Shotts Ch 7, Seeing the World as the Shell Sees It
  - Shotts Ch 8, Advanced Keyboard Tricks
  - Shotts Ch 19, Regular Expressions
- In-class activity

    - Clean up files from previous activities with
      mv $(ls –ignore=week*) week02
    - Batch Downloading and Building Simple Search Engines with Command Line Tools in Linux
- In-class discussion
    - How do we figure out whether a document is relevant to a query?
    - Why burst long documents into little pieces before putting them in a search engine?
- Linux commands to study and practice
    - **cat** and **echo**
    - brace expansion
    - command substitution
    - backslash escape sequences
    - **alias** and **unalias**
    - **clear**
    - **history**
    - tab completion
    - **wget**
    - **split**
    - **rename**
    - **swish-e**
        - Boolean Operators
        - Running Swish-e and Command Line Switches

Week 04: Named Entity Recognition

- Recap and Overview
    - We can automatically download arbitrarily large batches of files
    - We have a variety of techniques for analyzing text and finding patterns in it: word frequencies, concordances, regular expressions, search engines
    - A *named entity recognizer* is a program that goes through a text and tries to guess which words represent people, organizations, places or other kinds of entity
- Readings
    - Shotts Ch 12, A Gentle Introduction to **vi**
- In-class activity
    - Named Entity Recognition with Command Line Tools in Linux
- In-class discussion
    - Using probabilistic models for natural language

- Linux commands to study and practice
    - **vi**
    - **vimtutor**
    - Graphical vi/vim Cheatsheet and Tutorial

Week 05: Optical Character Recognition

- Recap and Overview
    - We have learned a variety of techniques for downloading texts, analyzing them and finding patterns in them
    - We can extract the (printed or typescript) text in photographs or digital scans of documents using OCR (optical character recognition)
    - Approximate regular expressions allow us to find terms that are *close* to a pattern, rather than matching it exactly
- Readings
    - Shotts Ch 17, Searching for Files
    - Shotts Ch 18, Archiving and Backup
    - Shotts Ch 20, Text Processing
- In-class activity
    - Doing OCR Using Command Line Tools in Linux
- In-class discussion
    - Taxonomy of digital sources, continued
        - born-digital vs. digitized
        - pictures of text (scanner, digital camera)
        - OCR: optical character recognition
    - What kind of errors are introduced by OCR?
    - How does approximate / fuzzy pattern matching work?
- Linux commands to study and practice
    - **locate** and **find**
    - **touch**
    - **xargs**
    - **gzip**, **gunzip**, **zcat** and **zless**
    - **zip** and **unzip**
    - **bzip2**
    - **tar**
    - **cut** and **paste**
    - **join**

    - **comm** and **diff**
    - **convert**, **display** and **identify** (imagemagick)
    - **tesseract**
    - **tre-agrep**

Week 06: Working with PDFs

- Recap and Overview
    - We have a wide variety of tools for working with texts
    - We can use optical character recognition (OCR) to extract printed or typeset text from digital images of documents
    - We can extract text, images, page images and full pages from PDFs with command line tools
- Readings
    - Shotts Ch 9, Permissions
    - Shotts Ch 10, Processes
- In-class activity
    - Working with PDFs Using Command Line Tools in Linux
- In-class discussion
    - How are PDFs structured?
- Linux commands to study and practice
    - **chmod**
    - **sudo**
    - **chown** and **chgrp**
    - **ps**
    - **kill**
    - **pdftk**

Week 07

- Recap and Overview
    - Foo
- Readings
    - Foo
- In-class activity
    - Foo
- In-class discussion
    - Foo

- Linux commands to study and practice
    - Foo

Week 08

- Recap and Overview
    - Foo
- Readings
    - Foo
- In-class activity
    - Foo
- In-class discussion
    - Foo
- Linux commands to study and practice
    - Foo

Week 09

- Recap and Overview
    - Foo
- Readings
    - Foo
- In-class activity
    - Foo
- In-class discussion
    - Foo
- Linux commands to study and practice
    - Foo

Week 10

Foo

Week 11

Foo

Week 12

Foo

Week 13

Foo