Text Analysis and Medical History

Ben Schmidt: NLM, April 13, 2016

Online notes: benschmidt.org/medhist16

1. (a) i. Outline
2. The Virtual Machine
   (a) Cutting and pasting.

   all programming is tweaking other people's code

   (b) Quick Start
3. Why Digital Text Analysis?
   (a) As a way of identifying important texts
   (b) For explorations, hypothesis generation, and sideways reading.
   (c) To expand audience for a set of texts.
   (d) The three operations of text analysis
      i. Choosing and understanding a set of texts
      ii. Defining smaller units of analysis: "words" and "texts"
      iii. Applying an algorithm
4. Selecting and getting to know a corpus.

   COHA: corpus.byu.edu/coha

   Careful Markup: Text Encoding Initiative (TEI)

   (a) You can analyze a textual corpus without doing text analysis!
   (b) Index Catalog
      i. Co-citation networks.
   (c) Google Ngrams (books.google.com/ngrams)
   (d) Where to get texts?
      i. General-purpose digital libraries.
   (e) Pertussis

- (f) Pertussis story
  - i. Medicine-Specific sources.

5. Defining Units of Analysis
   - (a) Optical Character Recognition
   - (b) Word Clouds
   - (c) Algorithms for tokenization
     - i. Named Entity Recognition

       Stanford Natural Language Toolkit

       - A. Part of Speech Tagging
       - B. Geo-parsing

         geocoding
   - (d) Defining Texts
6. Creating a corpus
   - (a) Regular expressions
7. Algorithms for insight
   - (a) For comparison
     - i. Addition, subtraction, division
     - ii. For comparison

       Odds ratio

       TF-IDF

       Dunning Log-Likelihood
   - (b) For Classification (“Supervised” machine learning)
     - i. Naive Bayes
   - (c) For Clustering (“Unsupervised” machine learning)
     - i. Principal Components Analysis
     - ii. Topic Modeling
8. Go-to-software packages:

   Cut and paste into an online environment: Voyant: voyant-tools.org

   Topic modeling and machine learning: MALLET: mallet.cs.umass.edu

Network Analysis: Gephi

Tutorials at ProgrammingHistorian.org

Cleaning and processing .txt files: Python

Statistical analysis: The “R” Language

Data visualization: R or D3

9. The Open Questions