Digital Text Analysis Workshop

Ben Schmidt: AHA, January 2, 2015

Online notes: benschmidt.org/AHA.pdf

1. 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
2. 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!
      i. Co-citation networks.
   (b) Google Ngrams (books.google.com/ngrams)
   (c) Where to get texts?
3. 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

4. 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
     - iii. Piping results into other sorts of analysis.
5. 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
6. The Open Questions