Starting to Text Mine the Digitized Library with HathiTrust Features

Ben Schmidt

2020-02-06

1. What books are scanned?
   (a) Who scans books?
   (b) Some ways of looking at what exists in Hathi
   (c) Everything you need to know about book scanning
      i. Book scanners
      ii. Stupid mistakes
      iii. OCR
2. How do you build up a list of Hathi volumes to work with?
   (a) Hathi Worksets
   (b) Keep lists of ids
   (c) Hathifiles (https://www.hathitrust.org/hathifiles)
3. "Feature Count" data.
   (a) Json data
   (b) Libraries

      Python: HTRC-Feature reader. (Organisciak et al). Powerful

      ```
      conda install htrc-feature-reader
      pip install git+git://github.com/massivetexts/htrc-feature-reader@caching
      ```

      R: "Hathidy." Schmidt.

      ```
      if (!require("remotes")) install.packages("remotes")
      remotes::install_github("HumanitiesDataAnalysis/hathidy")
      ```

      Javascript/Online. Suitable for quick looks.

      [https://observablehq.com/@bmschmidt/book-visualizations-sandbox]

[https://observablehq.com/@bmschmidt/some-notes-on-the-statistics-of-defining-characteristic-wo]

4. What sort of questions can you answer with these word counts, anyway?
   (a) Extending metadata and similarity
   (b) Exploratory
   (c) Substantive: Genre, etc.

   Underwood, Distant Horizons

   (d) Bibliographical