Notes-User Group Meeting-December 1, 2016

Date: December 1, 2016

Time: 2 pm EDT

Agenda: Extracted features dataset, presentation by Peter Organisciak

Notes:

Extracted Features Dataset (EFD) is a corpus of page-level extracted features (quantified counts) that can be used algorithmically in text analysis.

EFD is not contstantly updated, it is versioned. Available at analytics.hathitrust.org/datasets

Counts are provided for headers, footers, and body of the page. For every page the following counts are available: token counts (part-of-speech tags, case-sensitive), line counts, sentence counts, character counts.

Many possibilities exist with EFD: compare term counts, themes; identify parts of book; build topic models.

> *Example: Sam Franklin's project on creativity*
>
> *See also Peter's blog post https://porganized.com/2016/10/20/beyond-tokens-what-character-counts-say-about-a-page/*

Questions from the audience:

- How are duplications handled in the EFD?

  EFD counts features in all copies of the book, metadata can help to address that - linking duplicates. Future work.

- Will feature extraction code be available?

  Possibly, but the code is specific to how full text files are stored in HTRC, so the code is not generalized. It is available on github (github.com/htrc) and still needs to be generalized.

- What about OCR of non-Roman languages?

  Tokenization has been done for other languages, but counts do not exist yet. Japanese and Chinese need some work. Also, HTRC works with Google on re-OCR'ing some of the languages that were not done well. Will take time though.