Downie - HathiTrust

This page captures information about our support of Stephen Downie's faculty fellowship.

Overview

Title: Modeling the Massive HathiTrust Corpus: Creating Concept-based Representations of 15 Million Volumes

Link: http://www.ncsa.illinois.edu/about/fellows_awardees/modeling_the_massive_hathitrust_corpus_creating_concept_based_representatio

PI: J. Stephen Downie

Co-PIs: Peter Organisciak (HTRC, U. Denver); Boris Capitanu (HTRC); Craig Willis (NCSA)

In short, the fellowship is exploring the creation of reduced-dimensional term-topic matrices for the HathiTrust collection. This includes the exploration of scalable methods for dimension reduction/topic modeling (LSA/pLSA, LDA, autoencoders) for the full collection.

Updates

12/14/2017

- BW access finally in place as of 12/12, can start transfer process but need to enable Globus endpoint for HT data.
- Allocation will be used for two different projects related to HTRC – faculty fellowship and ngramming of HT data. Will meet with both projects teams on 12/15 to coordinate.

12/6/2017

- BW allocation approved, still waiting for access.
- Will work with Capitanu on sync'ing initial data for evaluation of deeplearning4j by end of week.
- Will meet with Co-PI Bhattacharyya 12/11 about BW project we are piggy-backing on

11/27/2017

- Conference call (Willis, Capitanu)
- Still waiting for BW allocation
- Boris explored deploying TensorFlow on TORQUE cluster and concluded that it's too complicated given that the deeplearning4j Spark already has a variational autoencoder implementation
- Will focus on deeplearning4j for now. Craig to request update on BW access.

11/20/2017

- Conference call (Willis, Capitanu)
- Discussed Tensorflow v deeplearning4j for scalable autoencoder implementations
    - Spark has support for SVD and LDA. Deeplearning4j add autoencoders for Spark.
    - Both can use GPUs
- Autoencoders
    - Proposing to use Sparse autoencoders
    - Hinton paper appears to be the motivation for applying autoencoders to text
    - Hinton and Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks
    - Lecture on youtube: https://www.youtube.com/watch?v=ARQ6PZh8vgE
    - Compare results to LSA only (on Reuters collection)
    - TensorFlow has VariationalAutoEncoder implementation as does deeplearning4j
- For next meeting, will prepare the following:
    - Shared access to either BW, ROGER, or IU (HTRC) cluster
    - Download and prepare Ted's 100K english volumes (need collection information)
    - Preliminary scaling of Tensorflow and deeplearning4j autoencoder with either Ted's or other collection
    - Access to BW allocation, if possible

11/17/2017

- Peter delivered sample autocoder implementation (set of Jupyter Notebooks)
    - https://github.com/craig-willis/htrc-autoencoder-examples
- BW allocation approved. Will need to send project information to initiate accounts.

11/13/2017

- Conference call (Willis, Capitanu)
- Spark has scalable SVD implementation (also LDA)
- IU has cluster with identical architecture to BW
- Ran original feature extraction code on BW
- Will review options and check in next week

11/12/2017

- Conference call (Downie, Organisciak, Willis)
- Most succesful autoencoder implementation is from Google/Tensorflow
- LSA may be possible at scale
- Peter has example running small set of cookbooks on HTRC server
    - Trimmed vocabulary
    - Might keep subset (e.g., ~4 million most common words)
    - Run over sub-batches of extracted features to determine which words are more useful
    - Should run at page (not volume) level

- Published extracted feature dataset
    - https://wiki.htrc.illinois.edu/display/COM/Extracted+Features+Dataset
    - https://analytics.hathitrust.org/datasets
        - Does not have extracted features
        - Extracted features
        - rsync just volume IDs
    - Get Peter's code running in that environment
- Will need to efficiently sample from those books
    - out of the box autoencoder code

11/7/2017

- Attended Faculty Fellows reception at NCSA.

9/17/2017

- Submitted BlueWaters proposal (in cooperation with Sayan Bhattacharyya and José Eduardo González)
- Title: Text Analysis of Books from the HathiTrust Digital Library to Characterize Descriptivity in Writing

7/31/2017

- Kick off
- By Aug 2 - 'To Read' list (Peter)
- This week - Peter and Boris touch base on initial workflow -> tractability of Peter's initial tinkering
- By Aug 11 - Wrap up lit review
- Aug 14 - Tentative meeting

References

- Geoffrey Hinton, Ruslan Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. Science, Volume 313, 2006.
- Ruslan Salakhutdinov, Geoffrey Hinton, Semantic hashing, In International Journal of Approximate Reasoning, Volume 50, Issue 7, 2009, Pages 969-978, ISSN 0888-613X, https://doi.org/10.1016/j.ijar.2008.11.006.
- Dipayn Dev. Deep Learning with Hadoop. http://proquest.safaribooksonline.com.proxy2.library.illinois.edu/book/databases/hadoop/9781787124769