LucidWorks Vectorize Module

May 7, 2013

CS 4624, Virginia Tech, Blacksburg, VA

Authors
David Kniphuisen, david21@vt.edu
Alan Tran, alan2438@vt.edu

Client
Kiran Chitturi, kiranch@vt.edu

Table of Contents

Abstract

The goal of our project was to create a learning module for students that are interested in converting a large amount of documents of data into a usable form for machine learning and searching. In order to complete this task, we wrote a module that gives information about how LucidWorks software handles this task of vectorizing a workflow. This module details the model that LucidWorks implements, as well as giving detailed instructions on how to create a collection, start the workflow, check the status of the workflow and finally access the results after the workflow.

Upon completion of our module, users will then be able to test on the example documents provided by the LucidWorks software, and be familiar with Hadoop's distributed file system. After users are familiar with how the software works, they are able to create their own vectorized documents. Our module also provides information about the installation of LucidWorks software on a virtual machine. However, if users have no access to the software, they will then be able to create their own instance of it by following the LucidWorks documentation for installation.

User's Manual

Users should start by watching either the powerpoint or video introduction to the module. They may then follow through each section of the module in the order that they appear.

Developer's Manual

Inventory of files:

- VectorizeModule.pdf
    - This file contains the module for how to run the vectorize workflow with the LucidWorks big data software.
- FinalReport.pdf
    - The final report detailing the deliverables and our project.
- ModuleIntro.ppt
    - A powerpoint that introduces users to our module
- ModuleIntro.mp4
    - A video where we introduce the vectorize module.

Lessons Learned

Timeline/Schedule:

- Get familiar with documentation (Feb 25)
- Finish creation of test data, and workflow (April 4)
- Complete body of knowledge for module (April 9)
- Finalize lesson plan (April 15)
- Submit for revisions to client (April 26)
- Finalize powerpoint and video, finish testing (May 2)
- Deliver project (May 7)

Problems:

The first problem we ran into during the creation of our module was the software availability. As there was a new version of the software being released when we were starting our work, we had to wait for the full new release to begin installation of the software.

The second issue we had was during the software installation and setup. There were errors during the installations process that delayed our work for a couple of weeks.

Solutions:

The first problem only took time, it was not a fixable problem. We simply had to wait for the new LucidWorks software to become available before proceeding.

The second problem also involved more waiting, as we needed our client Kiran to fix the bugs on his machine installation.

While both of these issues delayed the start of our work on the module itself, they did allow us more time to familiarize ourselves with the LucidWorks documentation.

Acknowledgements

Dr. Edward A. Fox, fox@vt.edu
Multimedia and Hypertext professor

Kiran Chitturi, chitturikiran15@gmail.com
Project Client

References

LucidWorks Big Data Vectorize documentation: (2012)
http://docs.lucidworks.com/display/bigdata/Vectorize

LucidWorks Installation documentation: (2012)
http://docs.lucidworks.com/display/bigdata/All-in-One+Virtual+Machine+Installation

Digital Libraries, LucidWorks Modules: (2013)
http://en.wikiversity.org/wiki/Curriculum_on_Digital_Libraries#Table_3._Modules_for_LucidWorks_Big_Data_Software

Vector Space Model: (2013)
http://en.wikipedia.org/wiki/Vector_space_model

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. New York: Cambridge University Press.