NLET National Laboratory for Education Transformation

October 16, 2018

Re: Sub-Award No X0089116-NP000

Sub-Award (SkillsCommons)

The National Laboratory for Education Transformation, NLET, received a DOL Sub-Award to work with SkillsCommons.org to increase the usability of the SkillsCommon repository.

Problem to be Solved (Aligning Education and Employment)

As the SkillsCommons repository grew its collection, the opportunity to make the repository more usable for deeper discovery and to match training materials to data about job postings and program courses, became apparent. TAACCCT grantees were uploading the Open Education Resources (OER) to SkillsCommons in differing formats and reliability. As a result, most community colleges turned in courses and programs within various Learning Management Systems "LMS" solutions (Blackboard, Canvas, Desire2Learnin, Moodle, etc.) and within different content solutions (Word, PDF, etc.). Each of these systems has restrictions regarding how they can be searched, data-mined, or interrogated. Consequently, more advanced search strategies could not be directly applied effectively to the current collection.

Objective of the Sub-Award (Discovery, Matching)

The objective of the sub-award was to increase the "discoverability" and "matching" capability within the repository. To carry out the objective, NLET initially conducted an extensive investigation to determine the ability to data mine within the various Learning Management Systems (LMS).

While there are industry standards (IMS) for the transfer of LMS content into what is known as the Common Cartridge data format, DOL had not required community colleges to move their content to the common cartridge. One LMS, Canvas (Instructure), uses the Common Cartridge as its native data format. Others such as Blackboard have the ability to convert their content. SkillsCommons has contracted with Blackboard to do this.

However, the detail and accuracy of discovery and matching rests on the ability to data mine the content by applying Natural Language Processing (NLP) and other Machine Learning (ML), and Artificial Intelligence (AI) techniques. These capabilities post-date the design of the LMS products. Also, for proprietary purposes, the LMS product companies protect the content in their systems. And each of these solutions has a license agreement that precludes their use by unlicensed recipients. Again, Canvas, which has an open option, and the open source platform, Moodle, are exceptions. Blackboard has such an option, but it is rarely used.

Because of this inability to easily search, scrape, or data mine in the LMS, and because there were multiple forms of content in each community college grant (such as descriptions, courses, reports, evaluations), more powerful methods were needed than those that are available commercially or within the open source communities.

In commercial content products, as commonly used in Web and app applications, the content is created in Content Management System (CMS) solutions that are routinely data mined. Anyone who is on and off the Web or using apps knows this is the case. Such content standards and methods are not routinely used in education because of proprietary concerns and older content designs.

Assembling New Methods to Mine SkillsCommons (Extraction)

NLET contracted via a sub-award with the Center for Data Science at UMass Amherst, engaging with a particular data scientist with extensive knowledge of data mining (NLP, ML, AI) but with limited knowledge of LMS solutions. The data scientist entered conversations with the SkillsCommons team, numerous community colleges that created courses housed in SkillsCommons, and technical and data teams at the LMS providers. For someone new to the field, this allowed NLET to understand the idiosyncrasies in skills and course data and how modern data techniques could be used for discovery and matches.

UMass extracted all the data and content within SkillsCommons to begin a process of indexing the contents and to further extract data from the LMS, PDF, and other content. (see UMass Summary)

Assembling New Methods to Mine SkillsCommons (Clustering & Alignment)

With the data extracted by UMass, NLET needed a technology partner that could conduct refined data discovery, alignment, and comprehension of the contents of SkillsCommons in order to take the content extracted by UMass and turn it into data that can be used in the following sample use cases of discovery and matching:

Sample Use Cases

- Course developer needs the best content to develop, update, or re-develop a course
- Employer needs to learn about a field or have trainees learn about a field
- Community college wants to look at open jobs serviced by a particular field or course
- Employer working with community college studies alignment with various courses.

NLET located www.Enlyton.com, an Austin-based technology company that works in deep discovery and alignment. Enlyton does not use a standard method of natural language processing to perform matches; instead they use a semantic data analysis powered by a math-heavy programming language to fine-tune alignment between different content databases. Their company does this in the fields of patent, medicine, and ecommerce.

Enlyton took the SkillsCommons data extracts from UMASS, applied their clustering solution, and then performed various experiments to make matches between:

- SkillsCommons content and DOL O*NET and SOC codes
- Between job descriptions (postings and O*NET and SOC descriptions)
- Between course listings and job postings
- Deeper discovery between one course or description and other courses.

An important aspect of the Enlyton solution is that any length of text, classification, or pertinent data can be used to power matches. Their solution is unlike other advanced search solutions with limited text inputs, or common natural language searches (NLP) techniques that can easily confuse search terms with multiple meanings (homonyms). The Enlyton solution builds the context for the terms and their usage. This allows very accurate alignment. (See Enlyton Summary)

Conclusions

The NLET sub-award provided methods and tests that proved that *the following are possible:*

- Higher levels of content discovery compared to advanced search
- Alignment between course content, classifications, resources, postings
- Viability of a repository beyond simple look-up of resources
- Research can lead to more integration between courses and jobs.

The final experiments that proved these points are included in attachments and can be performed for those who are interested. The limited size of this sub-award precluded developing user interfaces. The experiments are conducted at a code or text level.

Recommendations

NLET and the team strongly recommend the following:

- Creation and adoption of Open Occupational Course Standards (OOCS) and methods. Such course standards would avoid the problem of creating materials, resources, and courses in multiple and proprietary solutions that allow limited data mining or are under restricted commercial licenses.

- All content created for the workforce, an occupational, career, and technical skills markets whether open or proprietary, should inherently include the appropriate or approximate labor codes and classifications, or adopt a standard or schema that allows this to happen.

- The current SkillsCommons' default advanced search is optimized for most of those utilizing SkillsCommons. However, a new level service, informed by the NLET work, could be developed as a licensed or supported service for deeper discovery and matching

Gordon Freedman
President
National Laboratory for Education Transformation
225 Crossroads Blvd # 190
Carmel, CA 93924
www.NLET.org

N L E T National Laboratory for Education Transformation

UNIVERSITY OF MASSACHUSETTS AMHERST

College of Information and Computer Sciences

140 Governors Drive

Amherst, MA 01003-9264

CENTER FOR DATA SCIENCE

voice: 413.545.1323

fax: 413.545.1249

url: ds.cs.umass.edu

NLET / UMass / SkillsCommons Project Summary (10/9/2018)

Our exploration of SkillsCommons sought to gain an understanding of the corpus of uploaded courses, in terms of subject matter (topics, programs) and content metadata (course formats, types of files, etc.). In doing so, we hoped to find ways to improve SkillsCommons offerings, identify barriers to reusability, and find mechanisms to connect uploaded courses with applicable career paths and job openings.

The main project tasks were both qualitative and quantitative in nature:

- Conducted interviews with partner organizations (community colleges, academic researchers) to understand typical SkillsCommons use cases, how it could benefit from enhanced capabilities, etc.
- Created technical infrastructure for downloading and analyzing SkillsCommons course materials and metadata. The code created is open-source and posted in a public repository on Github.com. (provided on request)
- Exploratory data analysis on over 1,900 online and blended courses. Examined subject coverage, industry distribution, types of resources.
- Created advanced topic models over course materials using latent Dirichlet allocation techniques.
- Content extraction from over 150,000 course materials. Once extracted, data was assembled into a format usable by Enlyton tools.
- Produced several possible use cases for future iterations of the SkillsCommons platform.

Our main findings and recommendations centered around the fact that much of the data uploaded to SkillsCommons is hidden in archive files. By extracting available text and fully indexing it using Enlyton's tools, we enabled a much deeper course content search, as well as the ability to connect course offerings with appropriate industry ontologies and career opportunities.

Sincerely,

Matthew J. H. Rattigan
Center for Data Science
University of Massachusetts Amherst

NLET National Laboratory for Education Transformation

Enlyton

October 7, 2018

Attn: Gordon Freedman, NLET
RE: Skills Commons Project

Per the requirements requested, Enylton, Inc provided the following in support of the Skills Commons Project:

- Ingested content as extracted by UMASS for the courses within Skills Commons
- Crawled and indexed additional variations of the Skills Commons content as summarized in the attached overview
- Created and indexed a dataset of basic ONET job descriptions (SOC codes) as summarized in the attached spreadsheet
- Created a live demo for search interactions that produces cluster-based result output in real time based on a wide variety of user input queries.
- Ran experiments that utilize a job description as the search criteria to identify the most relevant courses within skills commons. Output can be produced in a demo user interface or exported to a spreadsheet file.
- Ran experiments that utilize a resume as the search criteria to identify the most relevant courses within skills commons. Output can be produced in a demo user interface or exported to a spreadsheet file.
- Produced a report where we evaluated each course summary within the Skills Commons repository (1831 courses) and semantically determined which of the 1110 ONET job summaries were most related/relevant to each course summary. The report lists which are the top 3-5 ONET job descriptions for each course.

Attached please find:

1. Description of Datasets, Indexes and search methodologies tested (SkillsCommons Discoverability.docx)
2. Spread sheet of Labor Code matches against each of the indexed Skills Commons Courses (Courses Tagged w-soc codes.xlsx)Spreadsheet of output for matched
3. Spreadsheet that highlights the ONET SOC code job descriptions used for matching (onet-soc-job-summaries-index.xlsx)

Kind Regards,

Chris McKinzie
Enlyton, Inc