Lakehead UNIVERSITY

Project 2017-17: Automating Gap Analysis of Learning Outcomes through Natural Language Processing

Final Report submitted to:
Ontario Council for Articulation and Transfer (ONCAT)
March, 2019

Acknowledgements

We would like to acknowledge and express appreciation for the efforts of software development team that include:

Dr. Vijay Mago (Project Lead)

Atish Pawar – Natural Language Processing algorithm

Sahib Singh Budhiraja, Jayant Chowdary – PDF Learning Outcome Extraction

Daniel Kivi, Joshua Romito – Web Development

Dillon Small – Data analysis and R+ support

Nancy Luckai – Strategic oversight

Kaylin Kainulainen – Kinesiology Content Expert

Table of Contents

Project Purpose and Goals

The purpose of this project was to develop an unsupervised Natural Language Processing algorithm and an online web-based interface that will automatically compare and contrast program and course level learning outcomes (LO) associated with post-secondary education. The goals associated with this project include:

Goal: To develop a Natural Language Processing (NLP) algorithm that identifies key aspects of syntax, grammar and semantics of post-secondary Learning Outcomes as informed by recognized taxonomies of learning.

Goal: To apply the NLP algorithm in a matching/gap analysis capacity whereby the algorithm is able to analyze two separate post-secondary programs and both accurately and reliably generate a list of matching Program Level and Course Level Learning Outcomes and a list of non-matching Program Level and Course Level Learning Outcomes.

Goal: To operationalize the Natural Language Processing tool such that it is applicable to large data sets (ex. Game Education Matrix) but also accessible to Ontario post-secondary institutions who have a database of learning outcomes and content data that could be used to generate transfer pathway possibilities. Ideally, if learning outcomes are associated with specific courses, the tool would recommend a list of courses that could be included in a block transfer credit package or that are duplicated in an integrated diploma/degree program.

Essentially, the goal was to create an online tool that post-secondary partners can use to upload their course information and learning outcomes into a database, select courses and run a comparison between either a program or a set of pre-selected courses. The NLP algorithm generates calculations of semantic similarity between learning outcomes, the courses associated with those learning outcomes and generates visual outputs that assist institutions in seeing the similarities and differences between their respective programs.

Methodology and Milestones

The project phases below describe the technical process our team used to develop an algorithm using Natural Language processes, the processes associated with the development of the web based application and concludes with a summary of all testing processes.

Defining Low and High Level NLP Tasks

The following phases describe the specific NLP related processes which may only be accessible to Computer Science specialists.

Pass 1: Maximize the similarity

The methodology used for this project considers the text as a sequence of words and deals with all the words in sentences separately according to their semantic and syntactic structure. The information content of the word is related to the frequency of the meaning of the word in a lexical database or a corpus. A semantic vector is formed for each sentence which contains the weight assigned to each word for every other word from the second sentence in comparison. This step also takes into account the information content of the word, for instance, word frequency from a standard corpus. Semantic similarity is calculated based on two semantic vectors. An order vector is formed for each sentence which considers the syntactic similarity between the sentences. Finally, semantic similarity is calculated based on semantic vectors and order vectors. Pass 1 deals with the three important aspects: Word similarity, Sentence similarity, and Word order similarity

Pass 2: Bound the similarity

The first pass of the algorithm returns the maximized similarity between two sentences. The second pass of the algorithm aims at computing a more robust similarity by reducing the ancillary similarity which causes skewness in results by considering syntactical structure, adjectives and adverbs, and negations in the sentences. Skewness in this context implies the deviation of the similarity from the similarity in the SICK dataset[1]. We use Spacy's dependency parser model which is the best performing model in the context of this algorithm. The intuitive idea behind this model is to keep track of the syntactical differences by incrementing a global dependency variable. The semantic analysis of any two sentences starts off with the comparison of words in the sentences and thereby determining the semantic similarity between all the words. Hence, the semantic similarity between words is the most crucial aspect when establishing the semantic similarity between sentences. Within the context of our testing, it was clear that the semantic relations between words in Post-Secondary Education are highly domain-specific. Therefore, we began to apply processes that would allow for the use of a domain-specific corpus to support enhanced word similarity.

---

[1] Pawar, A., & Mago, V. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. *arXiv preprint arXiv:1802.05667*.

Building a Domain Specific Corpus

Learning objectives from course outlines can contain peculiar words with meanings that are field specific. For instance, the word 'Python', in the domain of computer science, means 'A programming language' whereas it could mean 'A species of reptiles' in a more general sense. Hence, using a general purpose corpus does not always provide reliable results for highly specialized fields with technical jargon. So, building a domain-specific corpus and training the model with the corpus was applied in the later stages of this project. We chose Wikipedia as a source for compiling a corpus where a user can select a sub-category under which resides many sub-categories related to their domain to train the model. We used the petscan API to get the Wikipedia structure of a particular category, the Wikipedia python API to retrieve and parse the articles to get the textual content from the article webpage. Finally, we stored the corpus as a Python file which enables us to compile the corpus to find if there are any non-ascii characters. Filtering such characters is a necessary step before training the model. Every article is stored as a list element in the file for simpler iterations.

Developing the Online Database and Recommendation System

The purpose of the application is to give end users (Admissions/Registrars Office, Faculty, Chairs, Upper Administration and Students) the ability to upload course content and learning outcomes related to two different programs and receive a visual analysis of the similarities and differences. The following outlines the user experience and details associated with how the system applies various rules to sort through courses.

Uploading Course Content

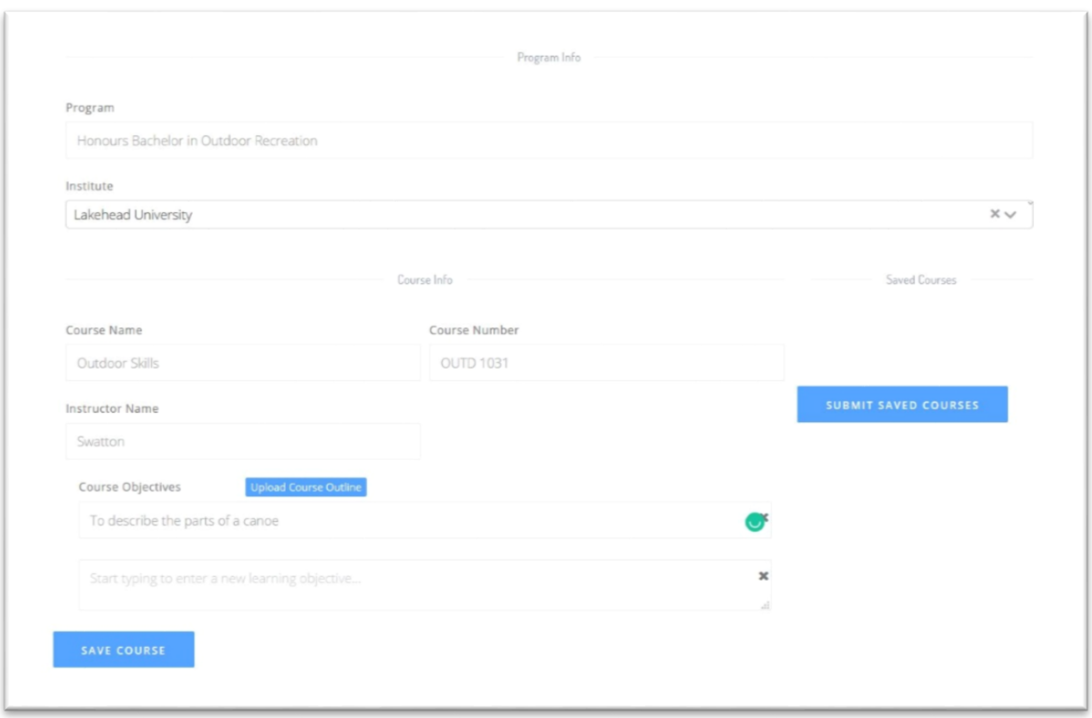

Figure 1

Figure 1 captures the web application upload screen. A user can input the name of the program, select from a list of all Ontario post-secondary institutions, input the course name and course code and the instructor (optional). Then the user can either upload a PDF of their course outline whereby the learning outcomes will be extracted or users can manually input their learning outcomes (entitled Course Objectives on the website). Users can then save as many courses as they would like and submit those saved courses into a database associated with the program and institution.

Accessing the Program Database

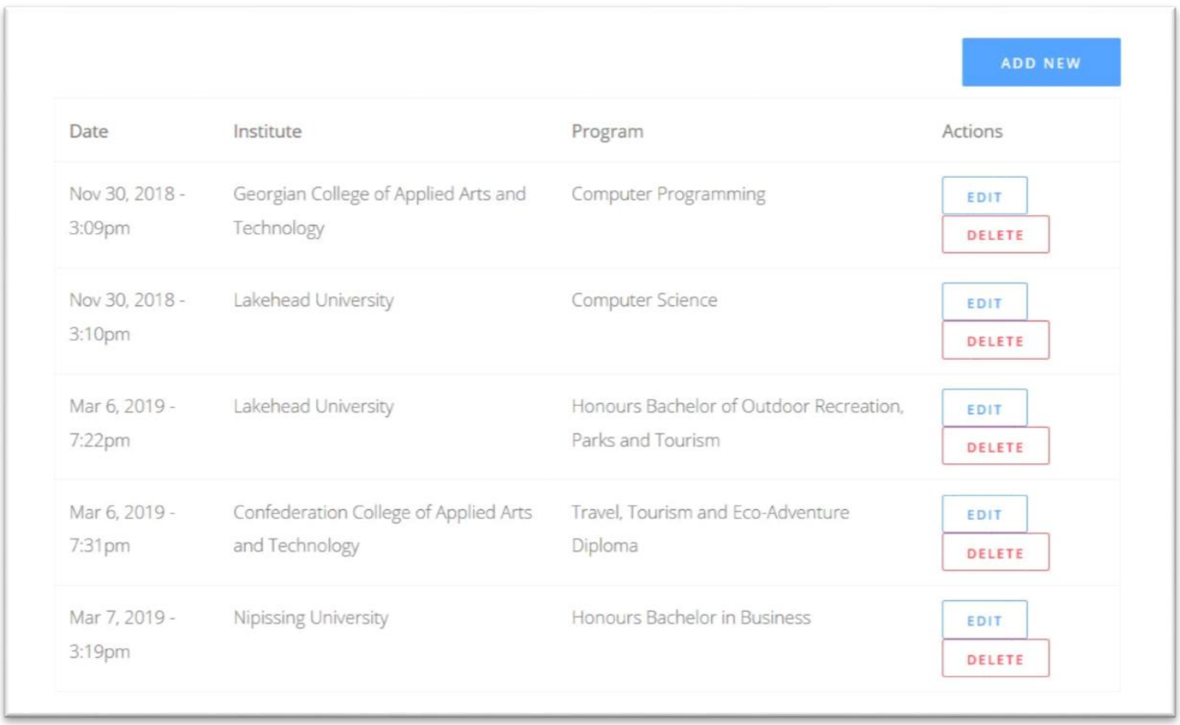

Figure 2

Once a user has uploaded a set of courses the user can access their database. This database contains all the courses and learning outcomes accessible to the user and the user can edit the courses or learning outcomes whenever they see fit. If this application was applied in the context of a transfer project initiative between multiple institutions shared access to the project data via the database could be provided.

This allows for an online accessible repository of all courses and learning outcomes associated with a pathway. All transfer related staff and faculty can view, update and change courses to modify agreements for program reviews, Ministry program standards changes, accreditation and cyclical program reviews.

Comparing Two Programs

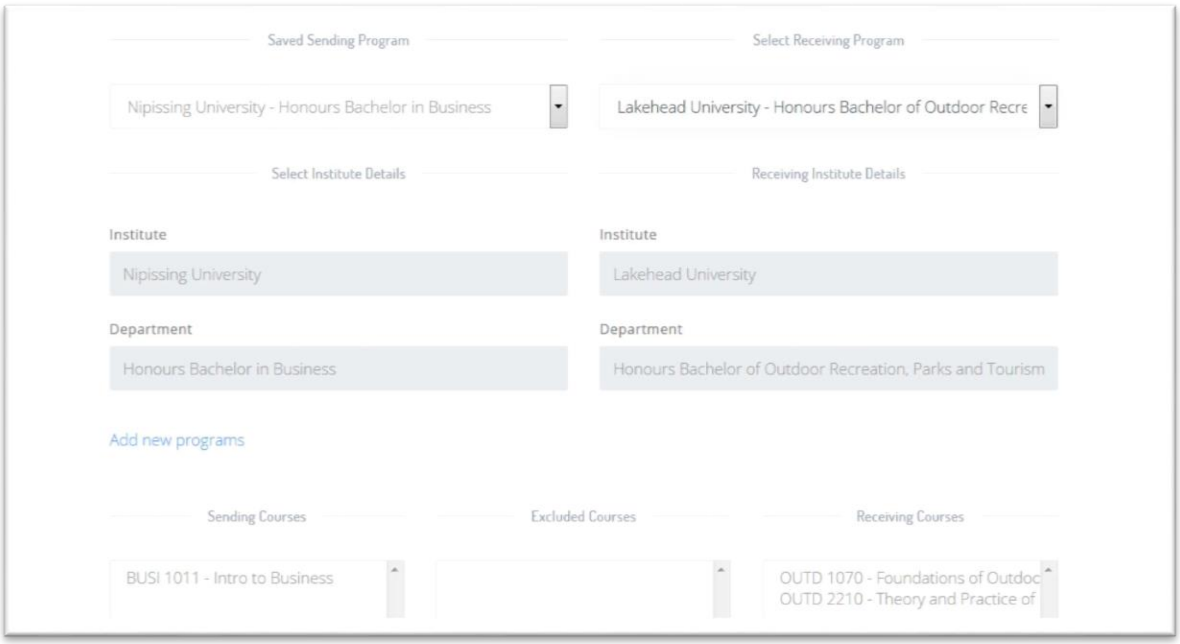

Figure 3

To compare the semantic similarities between two programs a user can select from their list of uploaded programs to identify a Sending program and a Receiving program. The page loads a list of courses associated with the selected credential and users can choose to exclude courses from the analysis. For instance, in one of our testing phases the user wanted to only compare anatomy courses between two health related credentials, therefore the user moved all unnecessary courses to the "Excluded Courses" box visible at the bottom portion of the image above.

Running an Analysis

Once the user has completed selecting the courses for analysis, they initiate the process by way of a "start processing" button and the system starts to compare every word of every learning outcome in every course and comes up a semantic similarity percentage between all sending and receiving institution learning outcomes. A score is calculated for each combination of receiving and sending courses by averaging all of the individual learning objective scores between the two respective courses. The learning objective score is simply the highest semantic match made between that learning objective, and any learning objective from the sending course. However, courses that have many semantically similar learning objectives should be preferred over courses that only have one of few similar learning objectives, so the learning objective with the highest semantic score is taken from the course with the most significant semantic scores, if such a course is available. The logic for these decisions is displayed in figure 4.

```
Input: Receiving and sending courses with learning
       objectives
Output: Semantic similarity value between 0 and 1
foreach learning objectives in receiving course do
    foreach learning objectives in sending courses do
        perform semantic analysis between all receiving
          learning objectives and sending learning
          objectives;
        best course ← course with most semantic
        matches 70% or higher;
        if best course is found then
            score for learning objective ← highest
            semantic match that belongs to the best
            course;
        else
            score for learning objective ← highest
            overall semantic match;
        end
    end
end
score for course ← average of learning objective scores;
```

Figure 4

User Output

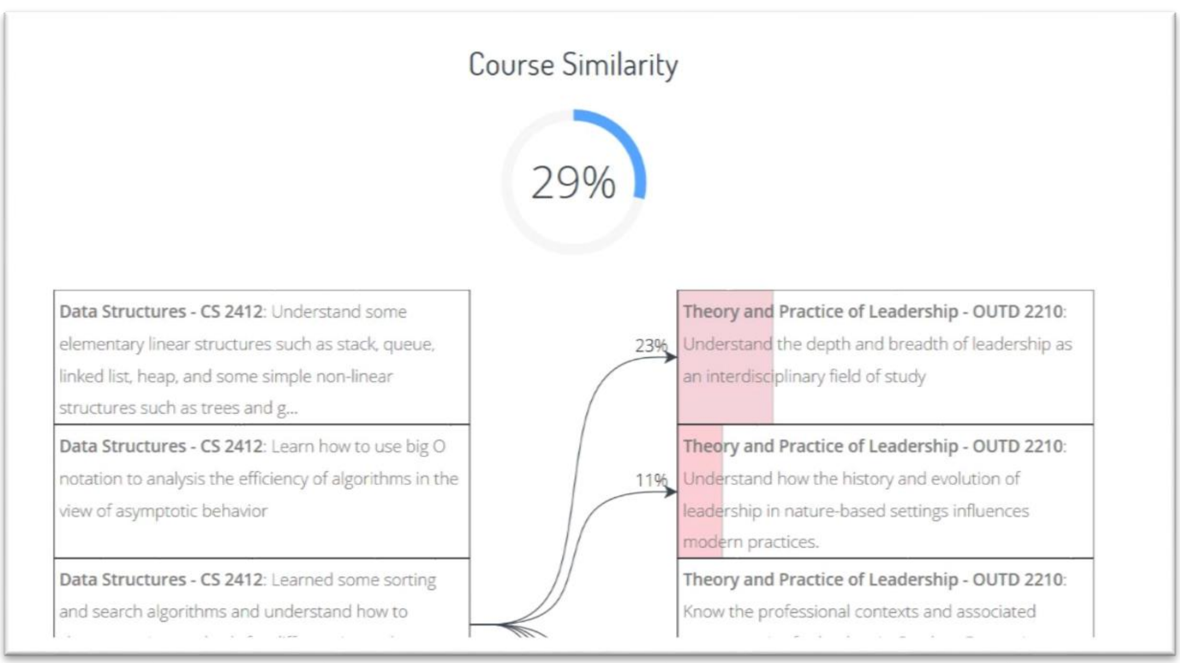

Figure 5

Users have three options to view the output from the NLP application analysis.

1. A results page, shown in Figure 5, that allows the user to select any set of courses from the overall analysis and see the exact semantic similarity calculation between any learning outcomes from those selected courses. An overall calculation of the similarity between ALL courses selected is displayed. In the example picture above the overall course match between the two courses selected is 29% (very low) with individual learning outcome percentages displayed below.
2. A report page can be generated that provides a:
    a. List of suggested courses for the user to consider for transfer credit that is ranked based on the highest percentage of semantic similarity.
    b. A heatmap of all the courses included in the analysis that allow users to see areas where different courses may overlap and see overarching trends for future investigation. In the example of a heatmap in Figure 6 the R series represents all the courses from one institution and the S series represents all the courses from a different institution. Note that the courses associated with the S and R series are not visible in the image.
3. An Excel file of all the learning outcome level and course level similarity percentages can also be generated and downloaded.

	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11
S1	53.19	56.67	36.83	76.41	72.61	63.96	76.98	77.43	76.03	60.14	69.51
S2	46.88	49.07	40.53	73.01	67.74	56.35	69.67	79.19	78.11	62.29	62.23
S3	55.35	48.85	40.31	74.95	72.11	54.44	74.99	76.16	82.14	69.72	63.45
S4	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S5	54.54	46.29	38.32	73.74	67.35	53.48	69.71	79.23	81.63	74.51	57.57
S6	48.35	56.67	39.75	77.04	71.47	61.37	77.27	72.09	74.83	56.80	70.26
S7	52.14	51.06	42.06	76.80	71.90	55.04	74.81	84.28	83.28	63.86	62.85
S8	49.28	47.14	41.25	75.31	67.44	55.37	70.96	77.25	78.24	67.94	60.18
S9	50.42	48.56	39.75	77.05	67.33	53.80	71.72	77.67	79.86	70.04	60.18

Figure 6

From the example heatmap above it is apparent that courses R7, R8 and R9 are more similar to the S series than courses R1, R2 and R3. Additionally, course S4 seems to have no similarity at all. The heatmap provides an overall visualization that can support transfer pathway discussions at a program LO level and insights as to where the overlaps exist between two programs.

Testing and Refining the Predictive Model

Algorithm Testing

The initial development of the unsupervised, general purpose algorithm included extracting the learning outcomes from 500 course outlines randomly selected and downloaded from a variety of global post-secondary institutions. Word to word and sentence to sentence comparisons were evaluated against standardized benchmarks in the field of Natural Language Processing including the SICK dataset and the Rubenstein and Goodenough benchmark[2].

Content Expert Comparison Testing

Course and learning outcome information from previous ONCAT funded pathways projects were used to test the accuracy and functionality of the system when compared to the decisions made by content experts and faculty members. We conducted three separate tests with historical data from ONCAT Pathway projects 2015-22, 2018-11, and 2016-24.

2015-22: Pathways to the Honours Bachelor in Outdoor Recreation

Of the multiple pathways developed in ONCAT Project 2015-22, we compared the historical data from the Georgian College Tourism – Marketing and Product Development diploma (17 course outlines containing 97 learning outcomes) and the Honours Bachelor in Outdoor Recreation degree (13 course outlines containing 101 learning outcomes). The analysis included approximately 20,806 sentence to sentence comparisons and took 31 hours and 45 minutes to process.

The style of gap analysis used by the project team in the historical project involved a 100+ page document containing tables where matching courses and learning outcomes were placed side and side with overlapping learning outcomes highlighted. Courses used in the historical project were shortlisted by the content experts and so we conducted our analysis on the short-listed courses.

---

[2] Pawar, A., & Mago, V. (2018). Calculating the similarity between words and sentences using a lexical database and corpus statistics. *arXiv preprint arXiv:1802.05667*.

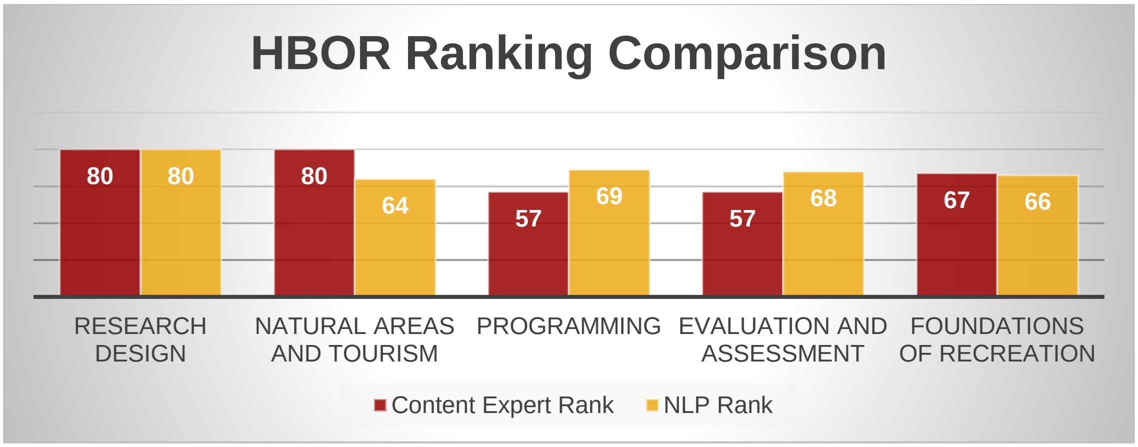

Figure 7

We elected to compare the percentage of overlapping learning outcomes between two courses selected by the content experts to the percentage of semantic similarity calculated by the NLP algorithm for courses that were included in the analysis.

Additionally, we were interested in whether the NLP system, as a recommender system where the highest overlapping course would be recommended first, would ‘choose’ the same courses and the content experts. For the HBOR pathways the NLP system would have recommended giving credit for Research Design, Natural Areas and Tourism, then Programming, Evaluation and Assessment and finally Foundations of Recreation.

The actual program credit awarded in the historical project included Foundations of Recreation and Programming. While the other credits were considered by the project team, they were not included for a variety of reasons beyond percentage of learning outcome overlap.

The outcomes from this initial round of testing include decisions to:

- Restructure the flow of database to decrease the amount of time needed to process the learning outcomes
- Noted that semantic similarity on highly technical outdoor courses was not a good match and we may require a corpus to train the algorithm in certain settings.

Brief Definition and Description of a ‘Corpus’: For the purposes of this report, a corpus is simply a large number of documents from a specific field of study. Used in a machine learning context, the system ‘reads’ all of the documents and builds a map of linkages between words to be able to calculate semantic similarity for a specific field. An example from this project would be the word Python. In the field of Computer Science Python is the name of a programming language, in general however a Python is a snake. When the word Python was compared to the word Language using two different general-purpose algorithms the measure of semantic similarity was 56% and 42% respectively. When a Computer Science corpus was used to train the system the similarity between the word Python and Language increased to 81% which is accurate only in the field of Computer Science.

- Future tests should either include a live content expert ranking or come from projects that include content experts ranking overlaps using surveys with percentages for a more accurate analysis.

2018-11: Two Way Transfer – Developing Post-Secondary Mobility Pathways for Ontario Health, Fitness and Well Being Students

To conduct an in person ranking of the individual semantic similarity calculations the project coordinator for project 2018-11 volunteered to do an analysis of the NLP output from learning outcomes associated with the anatomy courses in the Strength and Conditioning diploma from Canadore College and the anatomy courses required in the Honours Bachelor in Kinesiology at Lakehead University. After completing the analysis we downloaded the Excel files and the content expert ranked each learning outcome individually as either good, neutral or bad.

Out of 100 semantic similarity calculations between learning outcomes the content expert determined that 88 of the matches were ranked as bad, 6 matches were ranked as neutral and 6 matches were ranked as good. Considering the relative success of the Outdoor Recreation testing, this came as a surprise to the team. When we examined the language of the learning outcomes it became clear that for a highly specialized field with specific technical jargon, such as Anatomy and Physiology, it would be necessary to train the algorithm using domain specific databases (a corpus) of language.

Table 1 provides a sample of the Kinesiology output. The actual semantics of the sentence from the receiving institution anatomy course "Identify and describe the organization of the nervous system including anatomical and functional classifications" is similar to the sentence "Identify the structure and function of the circulatory system and some common disorders" however to any health professional these learning outcomes would be associated with two completely different aspects of human anatomy.

Table 1

Identify and describe the organization of the nervous system including anatomical and functional classifications.	**% Over lap**	**Course Number**
Identify the main components of the human skeletal structure and perform flexibility movements, including static, dynamic, and PNF.	82	REC 114 - General Exercise Protocol
Identify the structure and function of the circulatory system and some common disorders	78	REC112 - Anatomy and Physiology
Identify the structure and function of the respiratory system and some common disorders	78	REC112 - Anatomy and Physiology
Identify the structures, planes, regions, and structural levels of organization	77	REC112 - Anatomy and Physiology

The primary outcome from this round of testing was a determination to implement a corpus whereby a user could train the NLP system on the language specific to a highly technical field. If this application were to be applicable to fields as diverse as Engineering, Botany and Gerontology it was necessary to have a system that would recognize the language of each field and calculate semantic similarity accordingly.

2016-24: Honours Bachelor in Computer Science

For our third and final test the project team implemented a Computer Science corpus to analyze learning outcomes from a historical pathway development project. A component of the Computer Science pathways project included an online survey in which faculty members from both participating institutions ranked the percentage of overlap between a set of pre-determined courses via an online survey. Therefore, we could evaluate the difference between a human ranked percentage of overlap with an NLP algorithm ranked percentage of semantic similarity both on a scale of one to ten. Additionally, we processed all of the learning outcome and course comparisons using both the general purpose, unsupervised algorithm and also on a supervised algorithm trained to 'understand' Computer Science language using a corpus of specific Wikipedia documents.

To compile the corpus from Wikipedia 160,624 articles were collected from the sub-category 'Computing' using the Wikipedia python API. The articles were retrieved and parsed to extract the textual content and then stored as a Python file which is used to compile the body of articles and remove any non-ascii characters. Through this process our team recognized that we would be able to automate the process of developing a corpus using Wikipedia so that any user, including laypeople, would be able to select the appropriate category in Wikipedia and compile a domain specific corpus in real time. Completing this component could be considered for future, related work.

Table 2 and Figure 8 outline provide the results of this final test. The mean human similarity contains the average overlap of the indicated set of courses and respective learning outcomes using survey data. Six Computer Science faculty from Lakehead University and three Computer Programmer faculty from Georgian College completed the survey and ranked each of the 11 course combinations. The domain specific similarity column contains the results of the NLP system when the Computer Science corpus was applied to train the system. The General purpose similarity column contains the results of the unsupervised, general purpose algorithm with no field specific training.

Table 2

Institute 1	Institute 2	Mean Human Similarity	Domain-specific semantic similarity	General purpose semantic similarity algorithm
MATH1271	MATH1033	0.525	0.5804	0.18
CS1431	CS1008 + CS2006	0.775	0.7947	0.57
CS1411	CS1030 + CS2006	0.778	0.7830	0.53
CS2430	CS2125 + CS3025 + CS2068	0.775	0.7836	0.7
CS2477	CS1011	0.787	0.6707	0.66
CS2412	CS2021	0.7	0.6944	0.46
CS3412	CS3002	0.8	0.8133	0.49
CS4453	CS2018	0.5625	0.5852	0.9
CS4411	CS1045 + CS2068 + CS2070 + CS2099	0.3375	0.7846	0.85
CS4478	CS3023	0.65	0.7521	0.86
CS4467	CS3026	0.55	0.7958	0.39

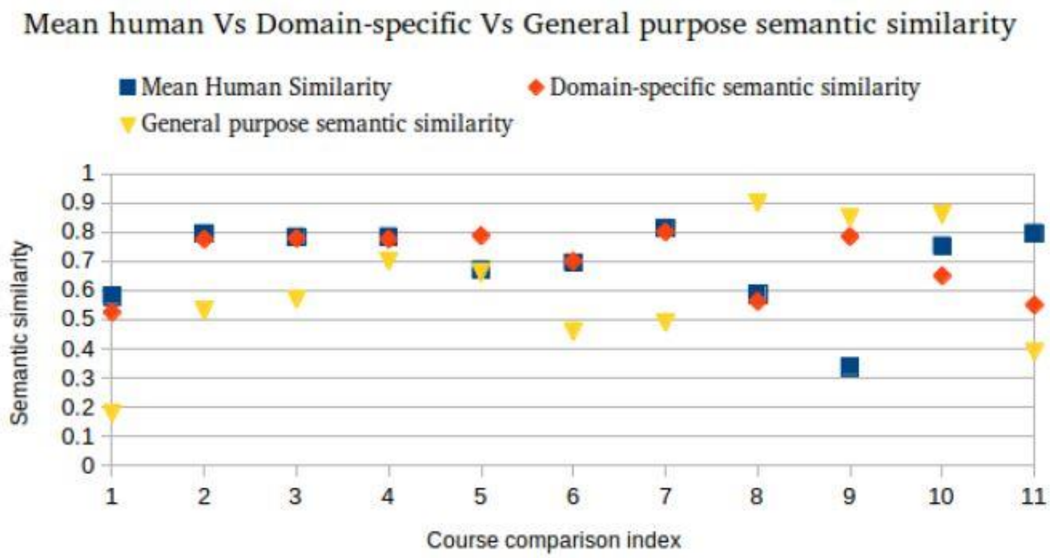

Figure 8

It is clear that the domain specific semantic similarity is closer to the content experts ranking of semantic similarity across all 11 courses with the exception of course number nine. This final test demonstrates that the NLP system can provide functional recommendations for both the overlap in course level learning outcomes but also in recommending courses to consider for transfer credit transfer pathway project teams. Figure 9 compares the top overlapping courses as ranked by all three testing systems.

Rank	Domain	Human	General
1	CS 3412	CS 3412	CS 4453
2	CS 4467	CS 2477	CS 4478
3	CS 1431	CS 1411	CS 4411
4	CS 4411	CS1431	CS 2430
5	CS 2430	CS 2430	CS 2477

Figure 9

Upon debriefing the results of this round of testing with the project team, all of whom are Computer Science experts and graduate students we came to following conclusions:

- Developing a domain specific corpus function is necessary for the accuracy and legitimacy of the NLP tool within the context of Post-Secondary Education (PSE),
- There are political and social factors that may influence content experts when ranking of similarity between two courses that could include:
  - Bias in the valuation of personal intellectual property with a bias to rank courses as unique and different from other courses with similar content
  - Perceptions that upper year level courses in a four-year program can not be equivalent to courses in a two-year program regardless of learning outcome overlaps

Conclusions

This project represents a significant step in the development of a professional and high-level application that implements machine learning to provide valid and reliable recommendations on the similarities between related courses across all the domains of the Ontario PSE system. To summarize, this project included the development of a general algorithm, a Wikipedia corpus compiling system template, an online transfer information database and accessible, user friendly course comparison system. While this system is currently functional and can be made accessible to stakeholders across Ontario, our team recognizes that more work can be done to:

- Increase the professionalism and user experience on the website
- Develop an easy and accessible function for users to select and compile a corpus that is specific to their domain.
- Incorporate improvements in the accuracy of the general-purpose unsupervised NLP algorithm