Mapping High-Impact Practices

By: Brendon Gu, James Pak, Leo Jung
Project Advisor: Zach Branson External Advisor: Joanna Dickert

Carnegie Mellon University
Dietrich College of Humanities and Social Sciences

Background & Introduction

High-Impact Practices (HIP) are courses that fall into one of several categories including:

- first-year seminars
- experiential learning
- capstone courses

High-impact practices require high levels of engagement and heavy interaction with faculty and peers, and are associated with increased student success.

The goals of our project are to:

(1) Effectively identify high-impact practices at CMU

(2) Map them to Dietrich College learning outcomes to improve the college's understanding of these experiences.

Data Preprocessing

Datasets and Relevant Columns :

Initial Course Information (22,107 rows; 15 features)	Dietrich High-Impact Courses (196 Manually Identified High-Impact Courses)	Additional Course Information (15,040 rows; 15 features)
• Course Department • Course Description • Course Number • Location	• Type of course (Capstone, Service Learning, First-Year Seminar)	• Course Units • Grad / Undergrad • Enrollment

Data Processing:

After joining the datasets and performing data cleaning, we used the resulting dataset to generate a document-term matrix and word2vec embedding to fit classification models with.

Data Cleaning	Document-Term Matrix	Word2Vec
1. Remove non-Pittsburgh, non-CMU courses 2. Convert words to lowercase 3. Remove whitespace 4. Remove duplicate courses and descriptions	• Rows: Course Descriptions • Columns: All the words that appear in the descriptions • Entries: Frequency of the word in each description	• Maps each word to a length 300 vector • Allows for a lower-dimension al representation of documents.

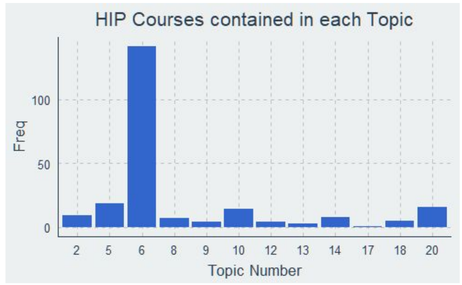
Methods

Unsupervised Learning: Classifying courses without true labels
- Manual Selection of Words, Document Similarity, Topic Modeling

(Semi)-Supervised Learning: Classifying courses with true labels
- Lasso Regression using Document Term Matrix, word embeddings

Course Classification

The structural topic model is a modification of the topic model that can incorporate additional document metadata. When fit on the dataset, almost ¾ of the high-impact courses were assigned to Topic 6.



For text classification, feature selection can reduce the dimension of the representation by eliminating useless features. A common method for feature selection is Lasso regression, which uses regularization to shrink coefficients to 0. We applied Lasso regression to the document-term matrix and word2vec embedding and compare their performance below.

Precision-Recall Curves

Accuracy

Document-Term Matrix	0.9419
Word2Vec Embeddings	0.9377

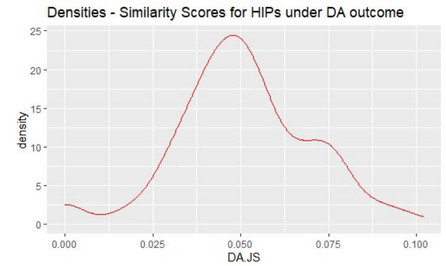
Computed using 5-fold cross-validation

Most Probable: 79-198 Research Training: History
"This course is part of a set of 100-level courses offered by Dietrich College departments as *independent studies* … In general, these courses are designed to give students some *real research experience* through work on a *faculty project* in ways that might stimulate and nurture subsequent interest in *research* participation…"
Probability: 96%

Learning Outcomes

We can measure the similarity between documentation for each of Dietrich College's Learning Outcomes and the descriptions of courses classified as HIP using the Jaccard similarity coefficient. In particular, we found capstone courses, independent studies, and undergraduate research courses to be most similar to the documentation for the Data Analysis learning outcome.

The distribution of similarity coefficients between HIP descriptions and the Data Analysis learning outcome documentation.



Conclusion

We had success applying both unsupervised and supervised methods to identify high-impact educational practices at CMU. We also made progress towards mapping courses to Dietrich College learning outcomes.

- Multiple unsupervised learning methods were tested but our attention turned to using supervised methods for classification after receiving labeled data.
- Supervised methods of using regression yielded better performance and were easier to evaluate and compare.
- Similarity is a promising way to link descriptions to the appropriate learning outcomes.

References

Kuh (2008), High-Impact Educational Practices
Mikolov et al. (2013), Efficient Estimation of Word Representations in Vector Space
Roberts et al. (2019), stm: An R Package for Structural Topic Models