# Engineering Vocabulary Development using an Automated Software Tool

**Mr. Chirag Variawa, University of Toronto**

Chirag Variawa is an accelerated-stream Ph.D. Candidate in the Department of Mechanical and Industrial Engineering at the University of Toronto. He earned his B.A.Sc. in Materials Science Engineering in 2009 from the same institution. He is the first Graduate Student member of the University of Toronto Governing Council elected from Engineering. His multi-disciplinary research uses principles from artificial intelligence, computational linguistics, higher education and aspects of neuroscience to investigate the design of engineering learning environments.

**Dr. Susan McCahan, University of Toronto**

Susan McCahan is a Professor in the Department of Mechanical and Industrial Engineering at the University of Toronto. She currently holds the position of Vice Dean, Undergraduate in the Faculty of Applied Science and Engineering. She received her B.S. (Mechanical Engineering) from Cornell University, and M.S. and Ph.D. (Mechanical Engineering) from Rensselaer Polytechnic Institute. She is a Fellow of the American Association for the Advancement of Science in recognition of contributions to engineering education has been the recipient of several major teaching and teaching leadership awards including the 3M National Teaching Fellowship and the Medal of Distinction in Engineering Education from Engineers Canada.

# Engineering Vocabulary Development using an Automated Software Tool


## Abstract

Understanding technical vocabulary is often a desired learning outcome in engineering education, and a significant part of professional communication in the engineering profession. Language used in engineering education plays a key role in creating an accessible and inclusive learning environment. The corpus of language common to both the instructor and student ought to converge as the student masters the course content. Instructors may currently use techniques to help identify this vocabulary, including referring to glossaries and increasing the frequency of their use in the classroom. There is an opportunity to increase transparency and accessibility to such vocabulary by developing an automated software-based tool that can be used by instructors to create customized course-specific wordlists for their courses. Using text extracted from instructional material in a course, the algorithm developed for this study is able to hierarchically identify and display course-specific terminology using principles from artificial intelligence, linguistics, higher education, and industrial engineering. Grounded in the theory of Universal Instructional Design, these wordlists can be integrated into a syllabus and then be used as a teaching aid to promote an accessible engineering education. The goal is to reduce barriers to learning by developing an explicitly-identified and robust list of vocabulary for all students in a given course. Creating an automated program that improves vocabulary information over time keeps it relevant and usable by instructors as well as students.

Presently, there is no automated method to develop course-specific vocabulary lists. To fill this gap, the authors have created a computer program, using a repository of over 2200 engineering exams since the year 2000 from the University of Toronto, which automatically identifies domain-specific terms on any given engineering exam. Specifically, each word from each exam is digitized and computed against others using a modified form of the Term-Frequency Inverse Document-Frequency (TF-IDF) algorithm to generate lists of context-specific characteristic terms. This well-known algorithm is used in the field of computational linguistics as a method of identifying words characteristic to a document, given a comparator set of documents. In this work, a modified approach has been developed that uses several comparator sets to produce a list of engineering vocabulary for a course. The effectiveness of this approach is evaluated by comparing the results to the judgment of subject-matter experts. This paper will use the data gathered to discuss the efficacy of this automated program in the context of engineering research methods, and will identify ways in which to make this program accessible to, and usable by, more educators in the field of engineering education.

## Introduction

This study investigates an approach to increase transparency of learning outcomes by explicitly defining them for students. Engineering students, in particular at the undergraduate level, are subject to understanding terminology relevant to their discipline, as well as the context in which these terms can be used appropriately. Through understanding of discipline-specific vocabulary, each student eventually forms a corpus of words that they can use as part of professional practice. As such, the importance of learning discipline-specific vocabulary forms a critical component of learning in engineering education, and is an area for research and optimization.

Currently, identifying discipline-specific vocabulary must be done manually. If the instructor chooses to, they will review course material and make a list of course vocabulary based on their subject-matter expertise. Sometimes, an instructor may defer to a "glossary" of a required course textbook, or the body of the textbook to support the teaching of vocabulary. However, this may imply that all terms are equivalently weighted in terms of importance; it relies on having an up-to-date text; and it relies on the text matching the instructor's terminology and teaching methods. In general, these manual processes are time-consuming and are not particularly rigorous to evolving knowledge and instructional environments.

An automated strategy can be based on existing instructional materials and be used as a starting point for further refinement by the instructor of the course. In this work we explore whether a computational method can be used to characterize vocabulary in engineering documents, and the efficacy of doing so. The approach used in this research is to develop and evaluate a computer program that can replicate human subject-matter expertise in characterizing vocabulary in instructional materials. This would provide a basis for further refining the learning outcomes to increase transparency, and as a result, accessibility to learning materials. The strategy for addressing this problem is to make design of vocabulary part of overall course design. This requires explicitly identifying the vocabulary that students need to learn in the course of their studies and is based on the framework of universal instructional design.

## Literature

This research is based on the framework of Universal Instructional Design (UID). The goal of the study is to increase accessibility to education by providing clearly-defined learning outcomes. In this specific study, this is done by identifying the discipline-specific requisite vocabulary that students need to master in a course. The UID framework is to "maximize accessibility to the greatest degree possible for the greatest number of users possible". Here, the research study attempts to maximize accessibility to language used in engineering education for students. As such, the principles of universal design should help guide research toward more accessible learning environment design for diverse student populations. There have been a number of authors who have interpreted the principles of universal instructional design.[1-3] The

universal design framework applies the principle of "learner centered" not just to one teaching instance, but to the design of the whole learning environment at every level. McGuire, Scott, and Shaw suggest that this framework is a "paradigm shift" that promotes uniformity of academic goals and standards by designing accessibility into a course, curriculum, and institution, rather than making exceptions for individual students who do not fit our preconceived idea of what is "typical".[1] They point out that individualized accommodation will still be necessary for some students. However, pervasive use of exceptions may undermine the integrity of a course, whereas designing accessibility into a course opens up learning opportunities for a broad range of students. Additionally, they have noted that this framework remains a largely untested strategy that requires further testing and validation. Pliner and Johnson discuss UID in relation to transforming social relationships which can be negatively affected by invisible barriers to inclusivity.[3] Their work suggests that implementing UID pedagogy creates a more "inclusive" environment which can decrease the barriers to learning that all individuals may have to some extent.

A review of the literature shows that there is serious concern about barriers to success for students, and a wide variety of approaches have been employed to try to mitigate barriers for at-risk students. Universal Instructional Design offers one possible approach and a framework for interpreting the impact of mitigation tactics. It will serve as a useful context for designing instructional tools that aim to maximize accessibility to education. However, instructors should also bear in mind that this is not the only framework and other ways of thinking about these issues should be investigated.

Building on the previous work performed by the authors[4], this paper expands on the modified Term Frequency-Inverse Document Frequency (TF-IDF) algorithm already used extensively in the area of vocabulary analysis. In summary, the TF-IDF algorithm is borne from the field of automated indexing and computational linguistics and a widely-accepted form of vocabulary characterization.[5-10] It takes an input document, stores each word into an array element, then performs a series of mathematical calculations to assign a numerical score to each word. This score is a diagnostic measure of how characteristic a word is to that specific document. This algorithm assigns this score based on term frequency, and how often each of the words in that document appears in a comparator set of documents. The TF-IDF algorithm is based on the following equation:

$$\mathrm{TFIDF} = \mathrm{TF} \times \mathrm{IDF}$$

where

$$\mathrm{TF} = \left(\frac{\#\ \text{of occurrences}}{\text{total}\ \#\ \text{of words}}\right)_{\text{in a single target document}}$$

And,

$$\text{IDF} = \log\left(\frac{\#\text{ of documents}}{\#\text{ of documents containing the word}}\right)_{\text{in a set of comparator documents}}$$

The TF-IDF equation is a measure of how characteristic a word is to a document, and can be discussed in terms of its constituent terms. The TF is a number determined by counting of occurrences of a particular word, and dividing that number by the total number of words in the target document: as such, it is a measure of frequency. The IDF is a measure of how important a particular term is within a set of documents, and is calculated by dividing the total number of documents by the number of documents in the set which contain that term, and then takes its logarithm. The TF-IDF formula multiplies these together and attaches the resulting score to each unique word in the target document. This equation works by comparing the static term frequency score for each word in an input document by a variable inverse-document frequency-score.

Since the comparator set of documents can change based on a number of factors, including year, instructors, etc., the IDF score can be updated and influence the TF-IDF score for all documents. This causes the TF-IDF statistic to evolve with changing datasets, and helps address the issue associated with evolving language. Additionally, the multiplication factor, the logarithm, enhances the effect of the document frequency and increases the resolution of finding characteristic terms within the input document. A high weight in TF-IDF is reached by a high TF and a low IDF of a word in the comparator set of documents. The weights therefore tend to filter out common terms. Since the ratio inside the log inverse DF is greater than or equal to 1, the value of IDF (and TF-IDF) is greater than or equal to 0. As a term appears in more documents, the ratio inside the logarithm approaches 1, bringing the IDF and TF-IDF closer to 0. This expands the effect of terms appearing in multiple documents, and maximizes its contribution to the TF-IDF score even though the TF score itself may be very similar to others in a particular document.

The modification to this approach, also discussed in previous work[4], is to use this TF-IDF algorithm repetitively in different contexts. Specifically, an input document can have the words within it be calculated using TF-IDF using one comparator set, and then calculated again using another comparator set. In both cases, the words will be the same since the same input document is being used. The TF-IDF scores, however, will be different because of the comparator sets. Based on the context being used, words will have a lower or higher TF-IDF score. Further, this phenomenon can be exploited to further extend the resolution TF-IDF scores, in particular by helping the experimenters discern vocabulary that is characteristic to an input document in a specific user-defined context – like a particular discipline, for example.

## Methodology

This study investigates the efficacy of the modified TF-IDF algorithm in mimicking human subject-matter expertise, as it develops wordlists of discipline-specific vocabulary. The methodology is comprised of two phases – the automated production of discipline-specific wordlists, and the testing of the efficacy of these wordlists. The first phase has been extensively published in previous work[4] and these results show that the TF-IDF algorithm appears to work. The second phase of the study, as discussed in this paper, uses subject-matter experts – faculty members –to evaluate the efficacy of the wordlists developed. The correlation between the judgment of the subject-matter experts and the list generated through the computational method is assessed.

The overall research study is outlined in Figures 1 and 2 below. Figure 1 shows phase one, and Figure 2 shows phase two, respectively. In phase one, words are prepared for analysis by converting all input documents to text-only format. Then the modified TF-IDF algorithm is used to develop word lists based on a target document (i.e. a specific document or set of documents from a specific course) and sets of comparator documents. The word list generated is a hierarchical discipline-specific vocabulary list that characterizes the target document. In phase two, human subject-matter experts were recruited to evaluate the efficacy of the automated approach in accurately identifying discipline-specific vocabulary.

The documents used for this study are 2254 electronically-available undergraduate engineering final exams from the University of Toronto. These exams are a summative assessment of a student's mastery of course concepts, and are intended to measure learning of the entire body of knowledge – or as close as possible – of a course. These documents are standardized across all engineering courses at the institution, are roughly the same length, administered in a closely-supervised environment, and are electronically available for data mining and study purposes. Due to the large quantity of words used in this study – over 22 million – this body of data serves as a starting point for additional research in the area of vocabulary characterization in engineering education.

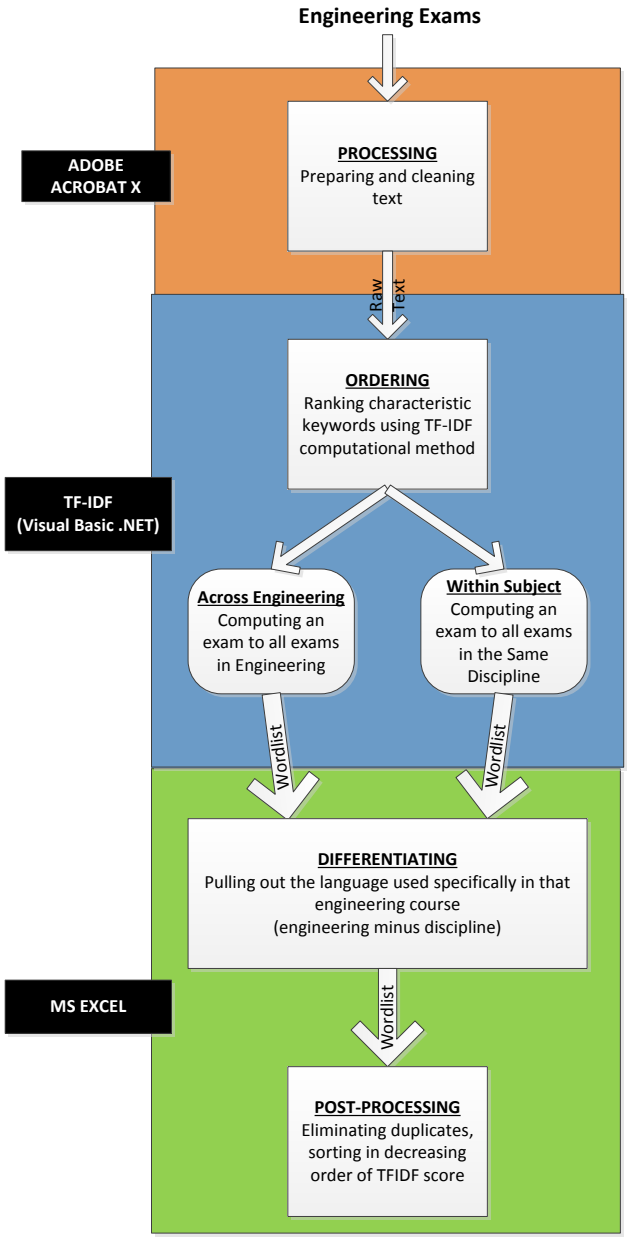


**Figure 1 - Shows graphically the methodology used in Phase One of the research study from top to bottom**

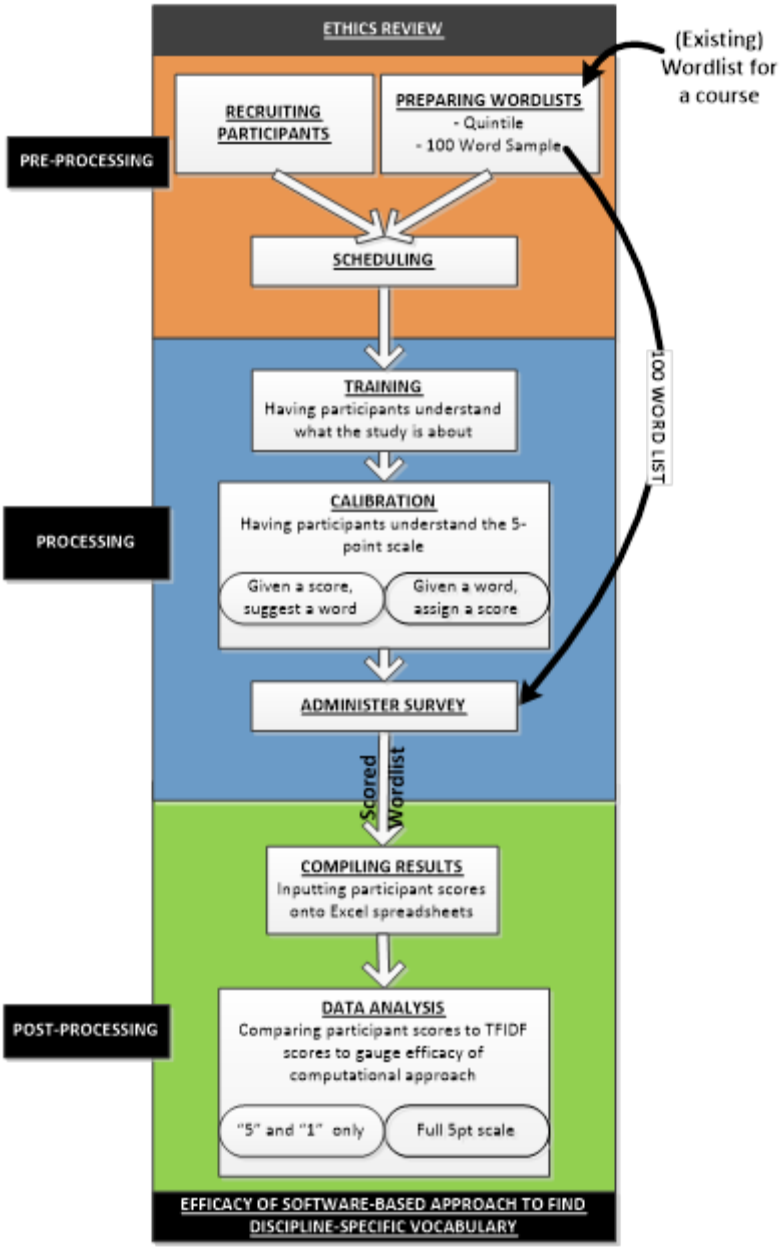


**Figure 2- Shows graphically the methodology used in Phase Two of the research study from top to bottom**

**Overview of Phase Two - Evaluating the Efficacy of the wordlists in capturing discipline-specific vocabulary**

This study focuses on gauging how well the wordlists capture discipline-specific vocabulary. To evaluate this, 9 subject-matter experts were recruited from the pool of faculty members teaching the courses whose exams were processed in phase one. As instructors, these faculty members are very familiar with the language that ought to be discipline-specific for the courses that they teach. This aspect of the research has passed the ethics review at the institution where this study was conducted.

The methodology of this phase of the research involves training, calibration, quantitative data collection, and debriefing of each participant. A condensed methodology is described below:

1. Participants were recruited using a standardized email request. In some cases, participants were asked in-person as a follow-up to the email, to ensure that the email was read.

2. A Doodle.com account was created, and each willing participant was scheduled into a 1-hour meeting timeslot; one participant per timeslot.

3. At each meeting, the participant was provided with an "Informed Consent" document. This required form was signed by each participant of the study. The study was briefly explained. This exercise reaffirmed the goal and purpose of this research, and emphasized the importance of providing authentic input.

4. The participant was told that they will be provided with a randomized list of 100 words, extracted from final exams of courses they have instructed in the past. Though the course for each participant was unique, each wordlist was developed using combined data across all years that the participant taught that course.

5. Participants were told that they would be assigning a number to each word in the list, using a scale provided to them. This is a 5-point scale, and ranged from words being not discipline-specific to very discipline-specific. A brief calibration exercise preceded data collection. The participant was given a print-out of the scale, and was given five words orally. The participant briefly discussed what they would score these words, and after they were confident in using the scale, the study progressed forward.

6. The participant was then given a list of 100-words from their own course and asked to assign a number from 1 to 5 to each word.

7. After completing the study, the participant was debriefed and given a complete wordlist for their course. This wordlist contained ranked words with corresponding TF-IDF scores, and a copy of a short academic paper explaining the study (written by the experimenter). Each participant was also thanked for their time and contribution to this study.

8. Each of the 1100+ datapoints (scores) were then manually entered into Excel spreadsheets for data analysis to measure how they compare to the TF-IDF generated wordlists.

**Results**

The results from this evaluation study are currently being investigated to understand statistical significance. Preliminary calculations show that the algorithm works well for a yes/no characterization – domain-specific or not-domain-specific – but is weak in identifying words that fall in between 2 and 4 (inclusive) along the 5-point scale. For example, the initial data shows that the program can identify words that are characteristic to a discipline or not characteristic to a discipline, but has difficulty in differentiating more finely words that are somewhat characteristic, as judged by the subject-matter experts.

A sample output that shows the TF-IDF output and the human subject-matter expert score is provided in Table 1. below. The full wordlist from a sample freshman electrical engineering final exam is condensed onto 100-words, and sorted in decreasing TF-IDF score in the left-most column. The word itself is in the second column, followed by its TF-IDF score. The participant-assigned score is a value assigned by the faculty member, and falls along a scale that ranges from 1-5(inclusive), with a high value indicating a high degree of confidence that the word is discipline-specific. The quintile-rank is a value determined by binning the 100-word sample wordlist into 5 bins, and is used to map the TF-IDF score for each exam to the 5-point scale used by the faculty members. As such, a quintile rank of 5 should correspond to a participant rank of 5, and so on for an ideal case.

**Table 1 –** Shows a condensed sample of a wordlist from a freshman electrical engineering final exam. The word list is separated into 5 quintiles, indicated by differences in cell colour, and ranked in decreasing order of TF-IDF score. For brevity, the 100-word list has been condensed to show a sample of words from each quintile. Correlations are highlighted in yellow to the right.

| RANK (/100) | WORD | TF-IDF SCORE | PARTICIPANT-ASSIGNED SCORE (/5) | QUINTILE-RANK (/5) | | |
|---|---|---|---|---|---|---|
| 1 | circuit | 0.033323128 | 5 | 5 | **100-word CORRELATION** (Using full 5-pt scale**):** | **0.7165** |
| 2 | voltage | 0.015487884 | 5 | 5 | | |
| 3 | electric | 0.014911103 | 5 | 5 | **100-word CORRELATION (**Using only extremes of 5-pt scale**):** | **0.9272** |
| 4 | capacitor | 0.009280436 | 5 | 5 | | |
| 5 | resistor | 0.00906219 | 5 | 5 | | |
| 40 | result | 0.000262347 | 3 | 4 | | |
| 41 | motor | 0.000260432 | 3 | 4 | | |
| 42 | discontinuous | 0.000254686 | 3 | 4 | | |
| 43 | tesla | 0.000239045 | 5 | 4 | | |
| 44 | deactivated | 0.000227847 | 3 | 4 | | |
| 70 | associated | 0.000121868 | 3 | 3 | | |
| 71 | respectively | 0.000121452 | 1 | 3 | | |
| 72 | half | 0.00011827 | 1 | 3 | | |
| 73 | results | 0.000117417 | 3 | 3 | | |
| 74 | losses | 0.000112727 | 4 | 3 | | |
| 81 | cannot | 2.31533E-05 | 1 | 2 | | |
| 82 | indicate | 2.30839E-05 | 3 | 2 | | |
| 83 | generated | 2.05447E-05 | 3 | 2 | | |
| 84 | difficulty | 2.03236E-05 | 1 | 2 | | |
| 85 | right | 1.88357E-05 | 1 | 2 | | |
| 91 | inside | -9.57969E-05 | 1 | 1 | | |
| 92 | variety | -0.000101296 | 1 | 1 | | |
| 93 | of | -0.000115615 | 1 | 1 | | |
| 94 | at | -0.000124816 | 1 | 1 | | |
| 95 | place | -0.000125485 | 1 | 1 | | |

**Discussion**

The data shows that a correlation exists between the participant-assigned scores and the software-assigned scores for the sample case chosen. An initial investigation of the data shows that an outright correlation across the full 5-point scale between software-scores and human-scores is present, but is not as high as a correlation between each of the extremes of the scale. In particular, the 5-point scale given to the participants maps to quintile categorization of TF-IDF scores. The sample case shows a correlation of 0.71, and this is similar to the other courses still being calculated for statistical significance. A preliminary observation of the participant-ranked scores suggests that though they have utilized the full-resolution of the 5-point scale, participants have a tendency of assigning very high or very low scores to each word. This appears to be consistent among all participants so far, and may suggest that a 5-point scale may have a resolution higher than what can be fully-utilized by each participant. Even though each participant was calibrated to the 5-point scale prior to beginning the study, favoring extremes on that scale might suggest that participants are not able to discern gradients in between and/or the extended resolution is too high.

If only the extremes of the scale are taken into consideration, the data shows that the computational method works very well. Specifically, if words that are scored a “5” or “1” by the participant are compared to their TF-IDF quintile bin, then there is a strong correlation. Sample data from a test case, a freshman Electrical Engineering core course, has a correlation of 0.927 and is shown in Table 1. Initial observations suggest that the subject-matter experts and the TF-IDF program are in agreement for high-ranked and low-ranked words, for most of the data collected so far. Currently, 11 studies have been completed, and 4 remain; the data so far suggest that the program works as the correlations are comparable across all of these courses.

When data is compiled from courses which may have less technical vocabulary, like design courses for example, an initial examination suggests that the correlations between subject-matter expert and the TF-IDF program are lower. In planning the survey, the experimenter predictively assigned three subject-matter experts to score the exact same design-heavy course. Though the data is currently being compiled, initial observations show that the correlation between participants and computer-assigned scores is much lower; slightly less than 0.7.

Currently a group of senior-year students from computer engineering are developing a web-based project based on the modified TF-IDF algorithm. The goal is to make this project accessible to people from around the world, so that they can submit their exams for calculation. This is in response to questions asked during ASEE-2013 where instructors wanted access to this software for their own courses. The users of this platform will have their documents categorized and added to the existing repository, and in return receive a scored wordlist based on the modified TF-IDF algorithm.

## Conclusions

The computational approach based on a modified TF-IDF algorithm appears to successfully replicate human subject-matter expert knowledge in identifying discipline-specific vocabulary. Through the dataset is currently limited to 9 exams, initial statistical measures for correlation show strong results. In particular, the software is able to accurately characterize vocabulary that is discipline-specific and this is a promising starting point for further research in the area of language analysis in engineering education. This work can lead to the development of clearer and more explicitly-defined learning outcomes, with the goal being to increase accessibility to technical terminology and robust vocabulary development for all students.

### References


1. Bowe, F., *Universal design in education: Teaching nontraditional students.* Bergin & Garvey, Westport, CT, 2000.

2. McGuire, J.M., Scott, S.S., and S. F. Shaw. Universal design and its applications in educational environments. *Remedial and Special Education,* 27(3), 2006, pp. 166-75.

3. Pliner, S.M., and J. R. Johnson. Historical, Theoretical, and Foundational principles of universal instructional design in higher education. *Equity & Excellence in Education,* 37(2), 2004, pp. 105-113.

4. C. Variawa, S. McCahan, and M. Chignell. “An Automated Approach for Finding Course-specific Vocabulary”. *Proc. of 120th ASEE Annual Conference and Exposition*. Atlanta, 2013.

5. Church, Kenneth W., and Robert L. Mercer. "Introduction to the special issue on computational linguistics using large corpora." *Computational Linguistics*19.1 (1993): 1-24.

6. McEnery, Tony, Andrew Wilson, and Geoff Barnbrook. "Corpus linguistics." *Computational Linguistics* 24.2 (2003).

7. Bybee, Joan L., and Paul Hopper, eds. *Frequency and the emergence of linguistic structure*. Vol. 45. John Benjamins Publishing Company, 2001.

8. SHI, Congying, Chaojun XU, and X. Yang. "Study of TFIDF algorithm." *Journal of Computer Applications* 29 (2009): 167-170.

9. Robertson, Stephen. "Understanding inverse document frequency: on theoretical arguments for IDF." *Journal of Documentation* 60.5 (2004): 503-520.

10. Singhal, Amit. "Modern information retrieval: A brief overview." *IEEE Data Engineering Bulletin* 24.4 (2001): 35-43.