Selection and Assignment of STEM Adjunct Faculty Using Text Data Mining

Andres Fortino, Qitong Zhong, Luke Yeh and Sijia Fang
agf249@nyu.edu, qz676@nyu.edu, cmy286@nyu.edu and scarlett.fang@nyu.edu

Abstract - **This paper presents the development and testing of a text data mining tool to assist in the selection and assignment of adjunct faculty to teach STEM courses. The tool scores the resume of a faculty member against course descriptions in a STEM graduate program. The tool returned a similarity score between a resume and course descriptions, which was then used as an indicator of faculty suitability to teach courses in the program. We enhanced the original tool with an improved user interface and deployed it to search for new faculty searches and in the process of assigning courses. A TD-IDF text analytic technique was used for similarity scoring. Our research question was to investigate whether a similarity-scoring tool for faculty resumes against course descriptions would be useful in the search and assignment process to hire faculty to teach specific courses. As part of our methods, we developed a friendly user interface to the existing tool using a student-centered coding contest. We applied the tool to the hiring and assignment of adjunct faculty. We measured success as the processing of a large number of open positions in a relatively short period of time and found a significantly high number of good fits between faculty and their course assignments. We investigated whether the scoring system positively correlated with the courses assigned to them. We successfully filled over 50 unassigned courses with appropriate faculty over a period of three months, where 30% were new hires. In the process, we discovered that the vast majority of the incumbent's similarity scores positively correlated to the courses assigned to them. This generated sufficient confidence that the description scoring system has been integrated as part of our faculty hiring and assignment processes in our programs.**

Index Terms - Similarity Score, text data mining, TF-IDF, resume, faculty hiring, STEM

INTRODUCTION

Program administrators and department chairs are often faced with the arduous task of finding appropriate adjunct faculty to staff courses in their programs. In some cases, the number of applicants, as well as the number of openings, is quite substantial. It is equally vital in the case of staffing a large number of course openings in the face of many adjunct candidates. Given that we have a faculty resume and a set of course descriptions as a stand-in for the job to be done, it is highly desirable to develop a simple tool that can be used a further filter in separating high potential candidates from those that, with a more time-consuming investigation, would not be appropriate. It is especially difficult in the case of STEM course offerings where, very often, the program administrator looking for technical experts to fill a highly specialized course, is not a subject matter expert. A tool that identifies and ranks the most likely candidates, no matter how basic, is a welcomed assist.

Our work is based on text data mining techniques previously applied to help students rank job openings against their program curriculum to look for the best fitting jobs [1]. We developed a Python-based text analytic tool that scores the resume of a prospective faculty member against the course descriptions in the STEM program. We tested the suitability of the program to filter a group of prospective faculty. We found that the scoring tool quite readily identified faculty who were suitable to teach a particular course. When we compared the selection of faculty without and with the use of the tool, there was a marked improvement in selectivity. The tool can also be used to rank many faculty resumes against one particular course description to filter and select the most appropriate candidates to teach the course. This can be used as a first-pass filter tool for hiring new faculty candidates. It may also be used to ran order incumbent faculty to select possible candidates to teach an unassigned course offering. The use of the tool does not replace the acumen of the hiring staff but gives them additional selection criteria to support their search.

LITERATURE REVIEW

Text data mining in higher education has gained much popularity, especially with great interest in curriculum analysis and design. Romero, in his survey of educational data mining [2], foreshadows the current high interest in text data mining applied to traditional education systems. He reported in 2007 on early efforts, which, ten years later, have progressed considerably.

Hsu and colleagues [3] developed a system for formative assessment of student learning in an e-learning environment by applying text data mining techniques to formative assessments. Using text analysis to create a mapping of job competencies to curricula was developed by Xun et al. at Singapore Management University [4]. Their procedure complements the work of Fortino using TD-IDF

and similarity scoring of text against O*NET job descriptions [1]. Unlike Xun, et al. [4], they did not take it as far as trying to extract core competencies to prepare for jobs that need to be covered by aspiring curricula.

Text data mining was employed by Debuse to study job postings in both Australia and the United States [5] and to derive key attributes of jobs in the ICT industry. These were then used to inform curriculum design. In our current work, we close the loop by tying curricula directly to job descriptions using similarity scoring of the text corpus. Data mining for insights from big data text corpus was used successfully by Liu and colleagues [6] to design a marketing analytics course in alignment with industry needs. Some workers, such as Nemeslaki [7], have used a reverse approach in their application of data mining. They start with a basic key concept or competency (in their case, it was "information security"), and it used to gather papers on the topic, which were then data-mined to extract core concepts that make up the field. Nemeslaki provides an intriguing approach that may be worthwhile to include in our future work. Similarly, text data mining was applied by West [8] in developing curricula for an interdisciplinary program in data science. They reinforced the designed interdisciplinary nature of the program by quantifying the coverage of design themes across the curriculum. There they first use it to develop the themes and assure they were covered in an interdisciplinary way [9] and then apply it to manage the teaching process by continuing the application of text analysis while the courses are running [8].

The application of text data mining to resumes to compare them to job descriptions follows two major approaches. The first and very successful approach is to extract key features from the resume to match it to job descriptions. The work of many researchers supports this approach [10], [11]. Some workers directly extract skills from resumes to give the recruiting staff needed information to match to jobs [12].

A second approach is the further processing of the text analytic results using alternative machine learning algorithms — ontological studies, sentiment analysis, or clustering based on similarity scores, for example. More recently, ontological approaches by Konys have yielded productive results [13]. Similarly, other text analytic techniques, such as factor analysis of the extracted features of a resume via text analysis, have been reported as yielding satisfactory results [14]. Verma creates a ranking index of resumes using cluster analysis of the extracted features. Other workers look for similarity between applicants based on similarity scoring of the applicants and subsequent clustering them based on those scores [15]. Although the use of resumes has been a recent focus for the application of text-analytic techniques, other sources of professional credentials and experience for a job seeker are now available. Online sources of a prospect's professional and work credentials and information, such as that available through LinkedIn, have been reported as the basis of text data mining to look for appropriate candidates [16], [17]. One intriguing development is the encoding of the judgment of recruiters (the subject matter experts of the jobs being filled) as the basis for machine learning [18].

Much of this work differs from our efforts in that they seek broad application to a wide range of resumes and typical job requirements. And they strive to be more precise and accurate in their matching. The work here is more simple: the use of the similarity scoring as an imperfect indicator as a filtering tool to reduce the task from many applicants to just a few that appear more appropriate as a first pass. This paper reports on the use of the most common credential tool, the resume, as the input data, and the course descriptions as the job descriptions. Exploring the benefits of the use of other possible, perhaps more robust, data sources (LinkedIn, curriculum vitae, syllabi) is left to future work.

RESEARCH QUESTION

Can text similarity-scoring tool that ranks a faculty resumes against course descriptions be a useful tool in the search for appropriate faculty and successfully assigning them to teach university courses?

Hypothesis

Relating to the primary objective of this research to test the efficacy of a literature search tool that improves the process of finding relevant literature for a literature review for research purposes by a novice two hypothesis were proposed:

Hypothesis 1: Using the text-similarity score to rank potential courses for a faculty to teach provides a significant improvement in the search and assignment process

Hypothesis 2: The text similarity score is an adequate indicator of a proper assignment of adjunct faculty to courses.

METHODS

Text Data Mining

A software tool was developed in Python based on earlier work [1] on text data mining for document comparison. It is identical to the tool developed for curriculum analysis and advising students for the appropriateness of jobs to their resumes as compared to their resumes. We had a number of text data mining techniques available to us, and as in the earlier work, we used Term Frequency–Inverse Document Frequency (TF–IDF).

TF-IDF is a process that converts documents into a numeric matrix. The idea of TF-IDF scoring is that when a word appears more often in a document than in others, this word represents more important information to this document. The way TF-IDF converts documents into a

matrix involves two calculations. TF refers to term frequency where IDF refers to Inverse Document Frequency.

TF is defined as the number of times a term appears in a document, divided by the total number of terms in the document. Therefore, the greater the TF means that, the more important the word is in that document. On the other hand, IDF measures how common a term is among all documents as the formula shown below, where *n* is the number of documents in the corpus, and DF means how many documents the word "I" appears at least once [19].

After transforming the resume and course description data into a latent space of lower dimensionality, the next step is to determine the similarity between a curriculum of the degree program and each of the course descriptions. Each course is converted to a query vector in the same two-dimensional semantic space that we chose to perform the SVD for the course text data matrix [20]. Then the cosine similarity is computed to measure the distance between a given query (the resume text) and the course description vector [21].

The Tool

The software tool was developed in Python using an extracurricular coding contest. Coding contest is an effective technique to provide students with a focused approach to learning informatics skills [22] and [23]. To develop Python coding skills, advanced training in Python, as well as a coding contest, was offered to students in the STEM graduate program. The students were challenged to create a simple to use interface to the earlier produced text data mining job-seeking tool. It was desirable that text data mining of prospect faculty's resume be easy to run by non-technical staff. The task involved further Python coding in building on a previously developed tool [1]. The winning entry used a Heroku-based solution [24], which is a free PaaS service, that enables the Python code to run in the cloud. Heroku provides all necessary environment to deploy and host the web application. The coding contest entries were judged by a development team of senior graduate students, recent program alumni, and program faculty. The winning entry became the search tool used in this experiment. The tool performs the following tasks:

1. Ingest course descriptions from a spreadsheet file as text data (the job descriptions). The web interface allows for this submission.
2. Ingest the resume or some other user-defined qualification document as a text file (UTF-8). The web interface allows for this submission.
3. Merge the files into an appropriate dataset and parse it into a TF-IDF representation.
4. Use a TF-IDF text data mining algorithm
5. Return a spreadsheet with a list of course descriptions scored against the qualification document rank-ordered by similarity score to the qualification text. The list should include the similarity score.

Using the Tool

The scoring tool requires two data files. The first is a simple (UFT-8) text data version of the faculty credential. For our experiment, we converted the resume received from the faculty candidate (the exemplar) into a UTF-8 text file. Any other text-based credential would do as well (e.g., a LinkedIn profile, curriculum vitae). The target file is a simple Excel flat file with course titles on one column and the text (UFT-8) course descriptions in the other. Additional information can be added in separate columns (course number, section, etc.), but the tool ignores all but the course description, and return these columns in the output document. Running the script (https://right-candidate.herokuapp.com/) and uploading the files as requested returns the course description file with the course scores and ranked by the score as well as an additional column with the similarity score as a .csv document, as well as displaying the results as a table on the web page. Users of the tool can then examine the top-ranked courses to confirm proper hiring and assignment decisions.

The tool can also be used in reverse. The course description as the exemplar and a set of faculty resumes as the target may be used to see who would be best suited to teach a particular course.

The Experiments

Experiment 1 - Hiring Effectiveness

For the preliminary experiment, our sample was new adjunct faculty hired and assigned to courses over a period of 18 months for a technical management graduate program at a major university. The intervention group (tool group) were new adjunct faculty hired in Fall 2019 The control group was new adjunct faculty hired in the Fall 2018 (n= 18) and Spring 2019 (n= 20) semesters without the use of the tool. The use of the tool was introduced in the hiring and assignment process for new adjunct faculty in the Fall of 2019 (n= 15). We controlled possible discrepancies in how they were hired by using the same recruiting process throughout the recruiting period. No changes in hiring administrators, faculty credential reviewers, hiring and assigning criteria, program curricula, or on-boarding process were made during that period.

In both cases (tool-use and non-tool use conditions), the need to hire appropriate faculty for courses was equivalent. Requirements to fill courses in the various subjects taught in the program did not substantially differ. No new courses were added to the curriculum, and no changes in the course syllabi were made.

The independent variable was the similarity score of each faculty's resume compared to the 37 course descriptions translated into ranks for each course among the 37 courses for that faculty member. The dependent variable was the likelihood that a faculty would be rehired to teach

that course. We also measure the number of courses in each rank category for each group. A first ranked course (r=1) had the highest similarity score.

Experiment 2 – Assignment Effectiveness

For the purposes of this second experiment, the population was also the group of current university faculty being reviewed for assignment to teach credit-bearing courses. Our sample was all currently active adjunct faculty (*n*= 98), assigned to teach courses in a graduate technical master's program at a major university over a period of 18 months, labeled as continuing faculty. We controlled possible discrepancies in how they were assigned by using the same search and assignment process throughout the assignment period. No changes in assigning administrators, use of faculty performance reviews, observation criteria for reassignment, program curricula, or faculty professional development programs, and curricula changes were made over that period of time.

The current resumes of the faculty were scored against the program course descriptions. We measured the similarity score of the course for each faculty and its rank within the whole curriculum (37 courses). We expected an adequate assignment if the course was in of the top 5 (by similarity score), or in the top 15% of the possible course assignments for that faculty.

RESULTS

Experiment 1 – Assignment Effectiveness

In the first experiment, we used the similarity scoring tool to improve assignment effectiveness. We used the tool to compute similarity scores and ranking of the 37 program courses to the resumes of all assigned adjunct faculty in the program (continuing and new). We did this for the three-semester period (Fall '18, Spring '19, and Fall '19). Figure 1 shows the distribution of a number of courses for each course rank category for all courses (n=218). Once the similarity scores of faculty's resume to program courses are computed and ranked, each course is then given a rank (r) for that faculty (r=1 for highest score). Figure 1 shows how many course assignments have been made to faculty by the course of their course rank of that assignment. Assignment of faculty whose course frank of that assignment in their top 5 ranked courses (r=1,2,3,4,5) account for 52% of all assignments. The top-ranked course (r=1) accounted for 25% of all assignments. Given that a human rater had assigned the faculty to those courses by evaluating the resume, the tool was able to support the rater had made a proper assignment. This supports the assertion that using the tool first, as a first-pass filter of resumes, makes the human rater's job easier by filtering the pool to the most appropriate candidates to teach a course.

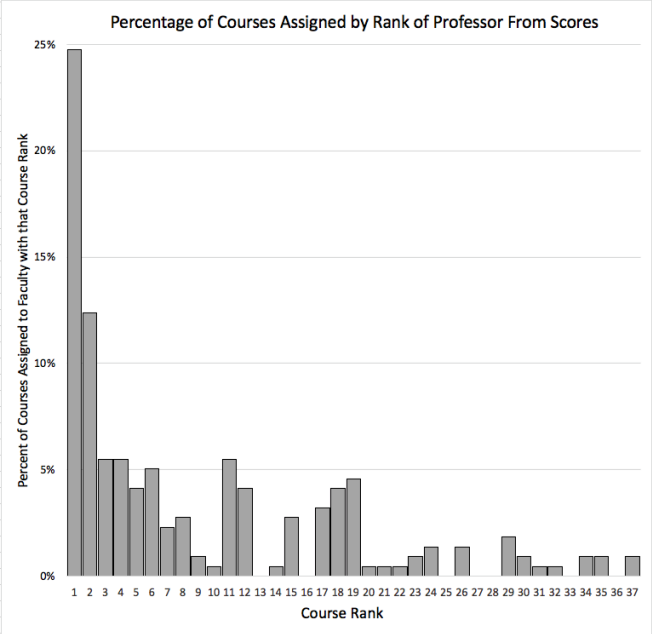

FIGURE 1
DISTRIBUTION OF COURSE RANK BY SIMILARITY SCORE FOR ALL COURSE ASSIGNMENTS.

Experiment 2 – Hiring Effectiveness

In the second experiment, we used the similarity scoring tool to improve hiring effectiveness. Figure 2 shows the distribution of similarity scores and course ranking among the top 10 ranked courses for all assigned adjunct faculty in the program over the 18-month period (Fall 18, Spring 19, Fall 19) broken down by courses assigned to continuing faculty (n=203) and newly hired faculty (FA19 only, n=15). The scoring tool was not used in the assignment of continuing faculty. It was used as a first-pass filter in making hiring and assignment for new hires for the Fall 2019 semester. There seems to be a visible shift towards using faculty with higher scores in the first three ranked courses for the new hires versus continuing faculty. It is not a statistically significant shift in using faculty with higher-ranked courses (χ^2=.9544, df=1, *p*-value=.3286). Fifty percent of the course assignments for new faculty was staffed with faculty whose scores placed the course assignment in the first of their three ranked courses. For the continuing faculty, 50% of the course assignments were filled with faculty whose course ranks were in the first five-ranked courses.

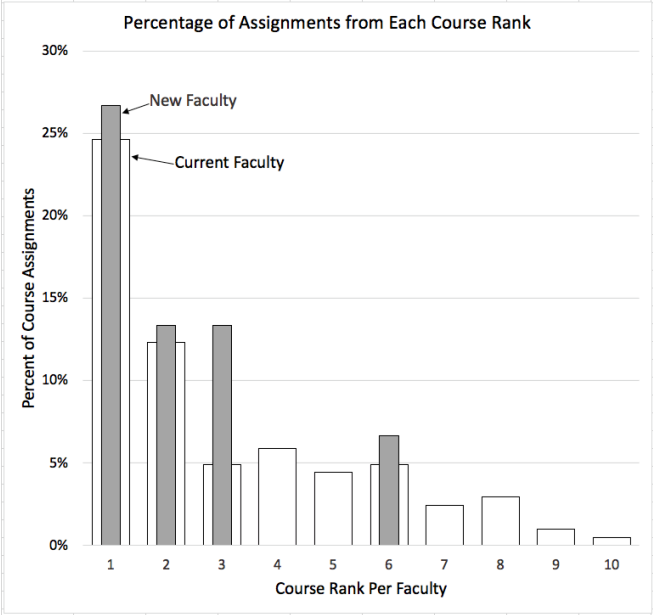

FIGURE 2
DISTRIBUTION OF COURSE RANK BY SIMILARITY SCORE FOR ALL COURSE ASSIGNMENTS.

Clustering Courses by Discipline

The tool was also used to ascertain if the assignments were in the right discipline for each faculty member. A TF-IDF similarity scoring analysis of the course descriptions against each other was performed. The resulting scoring matrix was used as an input to a hierarchical clustering analysis, which helped group similar courses into clusters by discipline. We used seven clusters as a reasonable grouping of the 37 courses in the program. Figure 3 shows the clusters by discipline and the number of courses in each discipline.

Course Cluster Name	Courses in the Cluster	Sections in FA18-FA19	
Operations Management	10	98	*Major Clusters*
Database	6	26	
Techology Management	6	32	
Management	5	34	
Finance and Risk	4	17	
Communications	3	9	
Systems Development	3	2	
Total	37	218	

FIGURE 3
CLUSTERS OF COURSES IN THE PROGRAM WITH THE NUMBER OF COURSES IN THE CLUSTER AND THE TOTAL NUMBER OF SECTIONS BY CLUSTER OFFERED IN FA18-FA19 SEMESTERS.

We consider the five course-groupings containing 84% of the courses in the curriculum as major clusters. For the purposes of this study, we consider a faculty with a course within the first three scoring ranks to have "expertise" in that course subject matter. Four of these major clusters were staffed by faculty with "expertise" in that subject. One cluster was completely staffed with first-ranked faculty (r=1 for that faculty). Figure 4 shows the staffing of courses in the major disciplines with their faculty course scores ranked in the first, second, or third ranks, and by percentage assignment by rank. The total number of courses in each discipline are shown as context to each discipline.

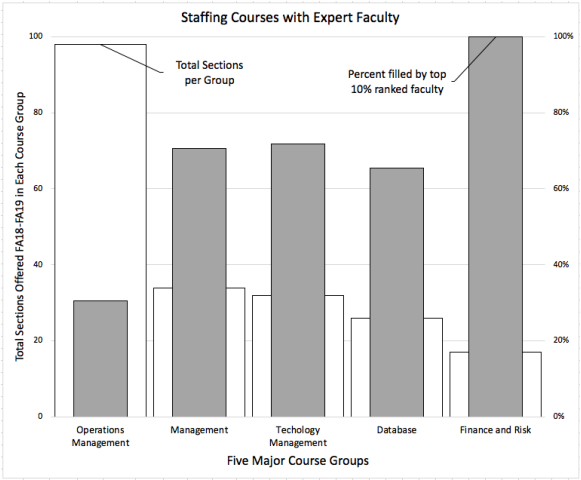

FIGURE 4
STAFFING OF COURSES WITH HIGHLY RANKED FACULTY AND AS A PERCENT OF ALL COURSES; ALSO BY THE NUMBER OF SECTIONS BY COURSE CLUSTER FOR THE MOST IMPORTANT COURSE GROUPS.

DISCUSSION

The results of this experiment are not definitive but provide a good indication of the usefulness of the tool. The results of experiment 2 showed that the null hypothesis may be rejected, and yet the tool does provide a good indicator of future faculty fit to courses in the program. Although not perfect, it provides an indication of subject matter familiarity and perhaps even "expertise" by the proposed faculty member. We can then use this tool to provide indications of faulty expertise match to a particular course and likelihood of success when assigned to that course and sue this tool as an appropriate first-pass filter.

The results of experiment 1 were more rewarding. The fact that so many of the faculty assignments were to courses where the faculty's resume similarity score places the course in a highly ranked position gave us encouragement to use this tool as a first-pass filter for assignments. There was a significant number of the total sections offered (10%) where the faculty resume score placed the assignment in the middle of the course rankings. We surmise that some courses are described in language that is seldom specifically used in a language that typically appears in a resume, especially by industry professionals. Such as courses such as Research Process and Methods or a Research Thesis capstone course emphasize topics not

common in industry resumes. The tool of more effective for courses with highly identifiable technical elements.

Figure 3 supports our contention that without the tool, the administrator and department chair must rely solely on their experience and acumen. With the tool, their job has become easier, and the time to identify appropriate faculty shortened. It is an imperfect tool, but as with many text data mining tools (such as sentiment analysis), a very useful adjunct to the decision making. The course staffing function is an arduous process for administrators without expertise, especially when they need to select faculty with deep subject matter expertise. The use of the tool reduces the effort by providing a first-pass filter - it considers an applicant for a course where their resume rank places them in the course in the top three ranks of all program courses.

Conversely, when looking at the pool of current faculty to fill an open appointment, the tool may be used as a filter to locate promising candidates as a first pass search. This works well, where there is a sizable pool of potential adjunct faculty candidates and dozens of open course sections.

The tool may also be used to search candidates for a given a course opening. This requires the scoring of all current and past faculty as well as any applicants and accumulating the results in a flat Excel database. An appropriate tabulating tool with filtering functions, such as Pivot Tables, may be used to obtain results quickly.

LIMITATIONS

We are not claiming this is a perfect tool or one that provides provable accurate results. For that reason, the tool should be used with care and then only as a first-pass filter to reduce from many choices to the few most promising ones. Also, resumes are not precise sources of information. The emphasis, the choice of words, the length of the document, make it hard to make exact comparisons from one resume to the next. As with most text analytic results, it is an approximation, a leading indicator at best. We are dubious of the use of this tool to compare one faculty's scores against another. Again, if it is to be done, it should be used sparingly and only as an indicator and not as an absolute metric.

Figure 5 shows an example of filtering the search process to fill a particularly challenging course section on Enterprise Risk Analysis and Mitigation, a highly technical course. Although the scores range widely in value, the fact that we filtered on course rank by similarity score to the top 2 ranked courses for each faculty yielded just a few potential candidates out of 100 faculty. Rather than selecting candidate B, who appears to have the highest score and ranking, we would investigate all five candidates further by examining their credentials more closely and using additional evaluation criteria. We have created a good starting point for further investigation for the appropriate faculty to fill the opening. It is interesting to note that for this course, no other candidates had rankings for this course above the seventh rank.

Faculty	First Ranked Course Score	Second Ranked Course Score
A		0.22
B	0.38	
C	0.22	
D	0.38	
E	0.21	
F		0.34

FIGURE 5
COMPARISON OF CANDIDATE RESUME SIMILARITY SCORES AGAINST THE ENTERPRISE RISK ANALYSIS AND MITIGATION COURSE DESCRIPTION AS A FILTER BY ONLY COMPARING CANDIDATES WITH SCORES IN THE TOP TWO RANKS FOR THE COURSE.

There are several limitations of this study, which we hope to explore and extend in future work. One question is whether the source document makes a significant difference (resume, full CV, LinkedIn profile.) Perhaps a combination or an aggregation of all three. Would the adding of a schedule of all courses taught in the past five years yield better results? On the other hand, perhaps a better "job" description would yield better results. Perhaps the use of the syllabus would yield a more suitable score. We hope to extend this work by investigating all these possible combinations to optimize the matching of faculty to a program and to course assignments.

CONCLUSIONS

We have shown that a simple text analytic tool may be built that improves the faculty recruiting and assignment process. Based on a basic recruiting input tool, the faculty resume, and the course descriptions as the job for comparison, a text analysis using TF-IDF similarity scoring helps rank and filter the right candidates. The tool is web-based and easy to set up and use, with no knowledge of Python required. With appropriate tabulating and filtering tools, quick searches may be conducted to discover adequate faculty to fill open sections, especially when the faculty pool and course sections to be staffed are sizable.

REFERENCES

[1] Fortino, A., Zhong, Q., Huang, W., Lowrance, R. Application of Text Data Mining To STEM Curriculum Selection and Development, IEEE ISEC'19 Conference, Princeton University, NJ, March, 2019.

[2] Romero, C., & Ventura, S. Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

[3] Hsu, Jung-Lung, Huey-Wen Chou, and Hsiu-Hua Chang. "EduMiner: Using text mining for automatic formative assessment." *Expert Systems with Applications* 38.4 (2011): 3431-3439.

[4] Xun, Law Sheng, Swapna Gottipati, and Venky Shankararaman. "Text-mining approach for verifying alignment of information systems curriculum with industry skills." In *2015 International Conference on Information Technology Based Higher Education and Training (ITHET)*, pp. 1-6. IEEE, 2015.

[5] Debuse, J., and M. Lawley. "Desirable ICT Graduate Attributes: Theory vs. Practice." *Journal of Information Systems Education* 20, no.

[6] Liu, Xia, and Alvin C. Burns. "Designing a marketing analytics course for the digital age." *Marketing Education Review* 28, no. 1 (2018): 28-40.

[7] Nemeslaki, A. Application of Science–Technology–Society Studies in Information Security Research, AARMS, Vol. 17, No. 1 (2018) 87–140.

[8] West, J. (2017). Validating curriculum development using text mining. *The Curriculum Journal*, *28*(3), 389-402.

[9] West, J. (2018). Teaching data science: an objective approach to curriculum validation. *Computer Science Education*, 1-22.

[10] Kudatarkar, V. R., Ramannavar, M., & Sidnal, D. N. S. An unstructured text analytics approach for qualitative evaluation of resumes. *International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN*, 2349-2163.

[11] Tripathi, P., Agarwal, R., & Vashishtha, T. Review of job recommender system using big data analytics. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 3773-3777). IEEE.

[12] Saxena, C. Enhancing Productivity of Recruitment Process Using Data mining & Text Mining Tools.

[13] Kony, A. An Approach for Ontology-Based Information Extraction System Selection and Evaluation. *Przegląd Elektrotechniczny*, *91*(11), 205-209.

[14] Verma, M. Cluster-based Ranking Index for Enhancing Recruitment Process using Text Mining and Machine Learning. *International Journal of Computer Applications*, *975*, 8887.

[15] Cabrera-Diego, L. A., Durette, B., Lafon, M., Torres-Moreno, J. M., & El-Bèze, M. How can we measure the similarity between résumés of selected candidates for a job?. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 99). The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp).

[16] Ha-Thuc, V., Venkataraman, G., Rodriguez, M., Sinha, S., Sundaram, S., & Guo, L. Personalized expertise search at LinkedIn. In *2015 IEEE International Conference on Big Data (Big Data)* (pp. 1238-1247). IEEE.

[17] Ha-Thuc, V., Xu, Y., Kanduri, S. P., Wu, X., Dialani, V., Yan, Y., ... & Sinha, S. Search by ideal candidates: Next generation of talent search at LinkedIn. In *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 195-198). International World Wide Web Conferences Steering Committee.

[18] Hardtke, D., Bollinger, J., Martin, B., & Vivas, E.. *U.S. Patent Application No. 13/662,312*.

[19] Kim, D., Seo, D., Cho, S., & Kang, P. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec. *Information Sciences*, *477*, 15-29

[20] Karakatsanis, I., AlKhader, W., MacCrory, F., Alibasic, A., Omar, M. A., Aung, Z., & Woon, W. L. Data mining approach to monitoring the requirements of the job market: A case study. *Information Systems*, *65*, 1-6.

[21] Huang, A. Similarity measures for text document clustering. In *Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand* (pp. 49-56).

[22] Burton, B. A. Informatics Olympiads: challenges in programming and algorithm design. In *Proceedings of the thirty-first Australasian conference on Computer Science-Volume 74* (pp. 9-13). Australian Computer Society, Inc..

[23] Skiena, S. S., & Revilla, M. A. Programming challenges: The programming contest training manual. *ACM SIGACT News*, *34*(3), 68-74.

[24] Middleton, N., & Schneeman, R. *Heroku: Up and Running: Effortless Application Deployment and Scaling*. O'Reilly Media, Inc.

AUTHOR INFORMATION

Andres Fortino is a Senior Member of IEEE. He is also the Chief Learning Officer at Autonomous Professional Development and a Clinical Assistant Professor of Management and Systems at New York University School of Professional Studies and Academic Community of Practice Leader in the Masters in Management and Systems program. He received his PhD in Electrical Engineering at the City University of New York. He is a member of the Academy of Management and INFORMS. His main area of research is evidence-based education, business analytics and data visualization, text data mining and its applications to higher education.

Qitong Zhong is a Senior Data Analyst, working in Omnicom group, and is an alumnus of New York University (2018). She obtained her Master of Science Degree in Management and Systems, with the specialization in database technologies and business intelligence. Her research interest is E-commerce consumer behaviors. She obtained her Bachelor degree in Sun Yat-sen University, Guangzhou, China. She published her undergraduate graduation thesis, Factors Affecting Trust Formation in the Context of Social Commerce, in a Chinese journal, Information Science.

Luke Yeh is a Master of Science Candidate in Management and systems at New York University. He is a data analyst with two years of experience in product operations and CRM in the technology and public sectors. During his master's studies, he served internships at the United Nations and in the New York City Fire Department. His research interest is the application of emerging technologies. He obtained his Bachelor of Science Degree in Transportation Technology and Logistic Management at National Chiao Tung University.

Sijia Fang was born in Luoyang, China. She is currently a graduate student at New York University's School of Professional Studies majoring in Management and Systems. She received her B.S degree in Financial Management from Shanghai University of Finance and Economics, Shanghai, China, in 2015. After graduation, she became a full-time auditor at KPMG China for two years, where she gained professional skills and insight into the finance-related area, which is experienced in qualitative and quantitative analysis and forecasting. She is proficient in Python, SQL, HTML, Tableau, and different data analysis and visualization tools.