Demonstrating Use of Natural Language Processing to Compare College of Engineering Mission Statements

Miss Sreyoshi Bhaduri, Virginia Tech

Sreyoshi Bhaduri is a PhD candidate at Virginia Tech Department of Engineering Education. She is a proponent for use of technology in the classroom as well as education research. Sreyoshi is a Mechanical Engineer by training, who likes programming to ”make life easier and efficient”. For her doctoral dissertation, she is exploring ways in which machine learning algorithms can be used by instructors in engineering classrooms.

Mr. Tamoghna Roy, Virginia Tech

Tamoghna Roy is a PhD candidate in the Bradley Department of Electrical and Computer Engineering at Virginia Tech. His research interests include statistical signal processing and applied machine learning. Tamoghna is currently working with the Hume Center for National Security and Technology on application of Deep Learning to Wireless Communication.

©American Society for Engineering Education, 2017

Demonstrating Use of Natural Language Processing to Compare College of Engineering Mission Statements

Most higher education institutions have a mission and/or vision statement that are designed to communicate with a variety of audiences. These statements are developed strategically by organizations and often reflect the college's unique vision which sets it apart from peer institutions. Analytical techniques which rely on word usage, semantic information, and metadata information can be used to generate powerful descriptive models with allow us to obtain relevant information from text-based data. This study presents a Natural Language Processing (NLP) based textual data analytical approach using Term Frequency-Inverse Document Frequency (tf-idf) to study the mission statements of engineering colleges/schools. A total of 59 engineering colleges/schools: 29 public, and 30 private, across the United States were analyzed in this study. Results of this study indicate that there is indeed a difference in tf-idf scores for public versus private engineering colleges. Tf-idf scores were computed for word tokens in the corpora, to go beyond frequency and capture relevance of a token for a given document. The contribution of this research is in the form of tables graphically summarizing the comparative word usage and providing a descriptive overview of the vocabulary of word tokens extracted from the statements analyzed. Topical grouping based on categories seen in existing literature was also conducted to analyze the comparative occurrence of tokens based on categories across the institutions. This study introduces a novel method for using NLP to analyze mission statements of colleges. Further, this study can help inform strategies on formation of mission and vision statements for universities by allowing administrators insight into vocabulary used across public and private colleges in the U.S.

Introduction

Large amount of text data is generated in engineering education on a regular basis. Analyzing these text data such as year-end reports, responses to open-ended survey questions, student blog posts or journal entries, and similar education data artefacts can yield interesting insight into student behavior, trends useful for learning analytics, and the like. However, analyzing these texts are time and resource intensive to inductively analyze large bodies data for individual instructors or researchers in educational contexts Variawa, McCahan, and Chignell (2013). Natural Language Processing (NLP) can help individuals interested in analyzing large quantities of textual data by processing data sets in both time and resource efficient ways.

One of the tasks that NLP is used for in analyzing textual data sets is that of stylometry. Stylometry can be understood as a linguistic analysis of use of words in terms of both choice and order, which can help characterize traits of a single or group of author(s). Fox, Ehmoda, and Charniak (2012) describe the underlying principle behind work on authorship attribution to be a set of statistically quantifiable characteristics of the writing style reflected by the word/phrase choices of individuals which make it easy to distinguish one author from the other. Stylometry is thus a type of quantitative or statistical analysis which helps identify and characterize specific literary or writing styles. In recent times, stylometry is used more popularly in conjunction with computational linguistic methods. For instance, in their study to analyze over 300 State of the Union Addresses, Savoy (2015) used machine learning techniques of clustering and NLP to

categorize authorship of the addresses, and to reveal trends in similarity and differences among styles. For example, they noted that President Obama predominantly used verbs in his speeches, indicating a speech oriented more towards action in contrast to President Hoover who used mainly nouns in speeches more oriented towards explaining situations such as the economic situation of the 1930s. Similarly, in another study conducted to explore the use of NLP to detect authentic suicide notes from elicited ones, Pestian, Nasrallah, Matykiewicz, Bennett, and Leenaars (2010) found that the machine learning algorithm outperformed the mental health professionals by correctly classifying 78% of the times (as opposed to 63% by the mental health professionals), whether or not a suicide note was genuine.

Thus, NLP can be effectively used to gather deeper insight from text based data on the characteristics or attribute (such as author, genuine or illicit, etc.) of the data. In this study we extend this concept of stylometry to understand the differences in the word use choices among private and public colleges of engineering in the United States. The purpose of this study is to use NLP based techniques to conduct an exploratory analysis of word choices in mission statements compared across public and private engineering colleges in the United States. The research questions driving this study are: 1. Which words are most frequently used in mission statements for private and public engineering colleges in the U.S.? 2. For six pre-determined categories of interest that each word may represent (eg. related to global, service, pedagogy, research, ethics, and diversity), what similarities and differences are noted in occurrences of each of the six categories across the two settings of public and private colleges?

Background

To help situate this research study, in this section, a discussion drawing from two bodies of literature is presented. The first part of this discussion focusses on prior work and existing research conducted in analyzing mission statements, with a focus on studies which emphasized on the differences between public and private colleges, and analyzed statements for engineering colleges. In the second part of this discussion the rationale for conducting an exploratory NLP based textual analytics, for this corpora of mission statements are presented. Thus the second section emphasizes the steady and emerging use of NLP techniques for use in education research.

Prior Analyses of Mission Statements

Mission statements can be described as public documents (Kreber & Mhina, 2007), readily accessible through the internet (Creamer & Ghoston, 2013), which are unique to institutions since they are intended to convey university-specific information reflecting the university's overall vision and purpose ((Kibuuka, 2001), as cited in Creamer and Ghoston (2013)), and are often developed through strategic planning in institutions. Thus, multiple research studies have acknowledged mission statements to be important in describing institutions intent and goals (e.g.,Tierney, 1999; Young, 2001), and have argued that institutions need to be more strategic in developing statements which truly reflect their characteristics (e.g., Barnett (2003) in Kreber and Mhina (2007)). In describing contradicting views on the significance of mission statements Kreber and Mhina (2007) cite Detomasi (1995) to describe how the latter suggest that mission

statements are “embarrassingly vague, and largely comprised of academic peties, dull platitudes, and odes of self-congratulation” (p.31) A similar sentiment is echoed in the work ofNewsom and Hayes (1991) who after analyzing mission statements for 114 US-based schools and colleges concluded them to be vague and lacking clear purpose.

Despite the contradicting views in existing literature on whether or not mission statements are truly reflective of the institution’s intent and goals, there has been ongoing research in understanding existing themes important to various institutions through analyzing mission statements both nationally and internationally (e.g.,Creamer & Ghoston, 2013; Kreber & Mhina, 2007). This can be attributed in part to the accreditation systems in place which demand that institutions have unique mission statements. Morphew and Hartley (2006) describe mission statements as being ubiquitous to institutions and reason “accreditation agencies demand them, strategic planning is predicted on their formulation, and virtually every college and university has one available for review”. Creamer and Ghoston (2013) acknowledge views by Delucchi (1997) and Morphew and Hartley (2006) both of who maintain that there is little research on mission statements in higher education, and add insight based on their search for such research by stating “a more accurate statement is that research literature is dispersed across a variety of disciplines…in fields as diverse as consumer studies, business, Christian education, and educational policy and management”. However, Creamer and Ghoston (2013) state that their search did not reveal any research on mission statements specific to engineering or STEM contexts.

The results of the mixed methods content analysis conducted by Creamer and Ghoston (2013) in analyzing 48 mission statements from engineering colleges serve to inform this research. Creamer and Ghoston (2013) mapped codes inductively derived from analyzing mission statements to the EC 2000 outcomes to test if any of the codes were associated.

NLP in education research

Bird, Klein, and Loper (2009) attribute the term Natural Language Processing (NLP) to describe any kind of computer based manipulation of a natural language. They distinguish between natural languages such as English, German or Hindi from what they connote artificial languages: such as those used in programming. They observe that the former have evolved over time, passing from generation to generation with rules that may be hard to pin down explicitly (pg.1). NLP can thus be understood as an interdisciplinary field in which computers are used to perform useful tasks involving human language (Jurafsky & Martin, 2007). Popular NLP tasks include that of: machine translation (e.g., Google Translate translating a phrase from Hindi to English), and automatic speech recognition (e.g., Siri on iPhones recognizing voice commands to “call home”), among others.

Through the use of NLP in educational contexts researchers have successfully demonstrated how NLP techniques can be useful to help educators and instructors gain deeper understanding of engineering education data sets for the purpose of assessment or to understand student vocabulary (for e.g., Magliano and Graesser (2012), Variawa et al. (2013)). Using NLP techniques in conjunction with the predictive capabilities of machine learning, Robinson,

Yeomans, Reich, Hulleman, and Gehlbach (2016) presented a model which learned from student responses and predicted course completion. In another research, extending computational linguistics through use of NLP for engineering education classrooms, Variawa et al. (2013) presented an automated method to develop course-specific vocabulary. Variawa et al. (2013) describe how their model successfully identified domain-specific terms on engineering exams using a modified tf-idf approach. In their paper, they also justify use of computational methods to characterize vocabulary in engineering documents for easy use by instructors, and further describe the efficacy of such methods. Thus there is increasing research using NLP and machine learning techniques to leverage data in educational contexts.

Methods

Data Collection

The researchers identified a total of 60 engineering colleges through random sampling from data publicly available by the American Society for Engineering Education in a book called the *Engineering College Profiles and Statistics Book.* Thirty of the colleges were public while the rest were private colleges of engineering. Following identification of the colleges, the researchers created an Excel database for the colleges wherein they assigned a column for mission statements (part of Excel file with Parent University names in Appendix A). The mission statements were extracted specifically from the engineering school or engineering college website. In one of the 30 schools identified for Public engineering college, no mission statement was found after a thorough search on the school's website. That college was excluded from the analysis, and a total of 59 engineering college/schools were analyzed.

Table 1: Describing the Proportions of the Public and Private Universities Analyzed

Control	Number (Proportion of Institutes)
Public	29 (49.126%)
Private	30 (50.814%)

Pre-Processing

Prior to analysis some pre-processing of the data was performed. At first all the punctuation marks are removed from the document and all the words are converted to the lower case. After that each of the word is lemmatized using Word Net (Fellbaum, 1998; Miller, 1995). Lemmatizing is a common technique used in NLP applications. A document can contain different forms of the same word (e.g. organize, organizing. organizes) which for analysis purpose should be considered as a single token. Lemmatization identifies the lemma of the word through morphological analysis. For example, after lemmatization, all the words in the above example (i.e., organize, organizing. organizes) will reduce to organize.

Feature Extraction

Tf-Idf features are used in this work for the subsequent analysis. In analyzing large corpora, tf-idf is a popular method which goes beyond the word frequency to not only compute frequency for the word in a particular document, but also multiplies this frequency by the inverse document

frequency which leads to lower tf-idf scores for words such as articles (a, an, the) which occur in large instances in both single documents and in the entire corpora (Church & Mercer, 1993; Shi, Xu, & Yang, 2009; Variawa et al., 2013). Variawa (2014) describes the computation of tf-idf scores:

> "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." (pg. 204)

In order to generate the features, the pre-processed documents are passed through the tf-idf feature extractor with unigram tokens. Many words in the document (e.g the, a, an, is) provide no information about the content. These are termed as stop-words and are disregarded by the tf-idf module. In addition to the common stop-words in English (provided by nltk library), 6 additional words were considered as stop-words – university, college, mission, student, school and engineering. The purpose of this work is to analyze the mission statements of engineering schools and compare them on the basis of whether the school is public or private. Since it is extremely likely that those 6 additional words will appear in all the mission statements irrespective if it is public or private those words are also included in the stop-words.

The output of the tf-idf feature extractor results in a matrix of dimensions 59x713 which implies that each document is encoded in a 713 dimensional feature space. Number of columns of the matrix (i.e. 713) also signifies the number of unique unigram tokens present in the corpus after the stop-words are disregarded.

Grouping Tokens into Categories

The categories inductively derived from analyzing mission statements for engineering colleges by Creamer and Ghoston (2013) were used to develop five categories of interest: Research, Service, Global, Diversity, Pedagogy. A sixth category for Ethics was also included. We then used WordNet (Fellbaum, 1998; Miller, 1995) which is an open source database that contains words, their root words and related synonyms to develop a list of "tokens of interest". The corpus was then searched for these words and their frequencies. Analysis described in the Results section elaborates how this grouping of tokens was used for comparison across the sets of Public and Private colleges.

Table 2: Categories of Interest Developed Based on Existing Literature

Categories	Synonyms Generated Through WordNet
Pedagogy	knowledge, apply, math, science, cognition, noesis, use, utilize, utilise, employ, hold, go for, put on, lend oneself, give, practice, enforce, implement, mathematics, math, scientific discipline, skill
Research	research, design, experiment, innovation, real, publication, publish, innovate, innovator, inquiry, enquiry, search, explore, design, plan, blueprint, pattern, figure, purpose, intent, intention, aim, invention, excogitation, conception, project, contrive, experimentation, try out, initiation, founding, foundation, institution, origination, creation, introduction, instauration, real number, existent, tangible, actual, genuine, literal, substantial, material, veridical, very, really, rattle, 'issue', publishing, print, bring out, put out, release, write, introduce, pioneer, trailblazer, groundbreaker
Ethics	ethic, society, environment, ethical motive, moral, morality, moral philosophy, ethic, moral principle, value system, value orientation, ethical code, club, social club, guild, gild, lodge, order, company, companionship, fellowship, high society, beau monde, smart set, bon ton, environs, surroundings, surround
Global	global, world, planetary, worldwide, worldwide, ballshaped, globose, globular, orbicular, spheric, spherical, universe, existence, creation, cosmos, macrocosm, domain, reality, earth, globe, populace, public, worldly concern, earthly concern, human race, humanity, humankind, human being, human, mankind, man
Service	service, outreach, religious service, divine service, military service, armed service, service, robert william service, avail, help, table service, service, serve, serve, service of process, overhaul, inspection and repair
Diversity	diversity, woman, minority, interdisciplinary, diverseness, multifariousness, variety, woman, adult female, charwoman, char, cleaning woman, cleaning lady, womanhood, fair sex, nonage

Results

Our findings suggest that the mission statements for the colleges of engineering are different for private and public colleges. In this section the results of analysis for the comparison are presented. Specifically, the research questions driving this study are answered: 1. Which words are most frequently used in mission statements for private and public engineering colleges in the U.S.? 2. For six pre-determined categories of interest that each word may represent (eg. related to global, service, pedagogy, research, ethics, and diversity), what similarities and differences are noted across public and private colleges?

1. Frequently Used Words

In order to determine which word tokens were most relevant as well as frequently used in each set of mission statements (ie. from public or private engineering schools/colleges), we first determined the tf-idf scores for each word or token. Thus, post the pre-processing to remove stop words, and lemmatize the raw words, in the data unigram tokens or singular words were extracted. For each token a corresponding tf-idf score was assigned. The tf-idf score, as explained earlier goes beyond the frequency to assign a score which is more suitable to

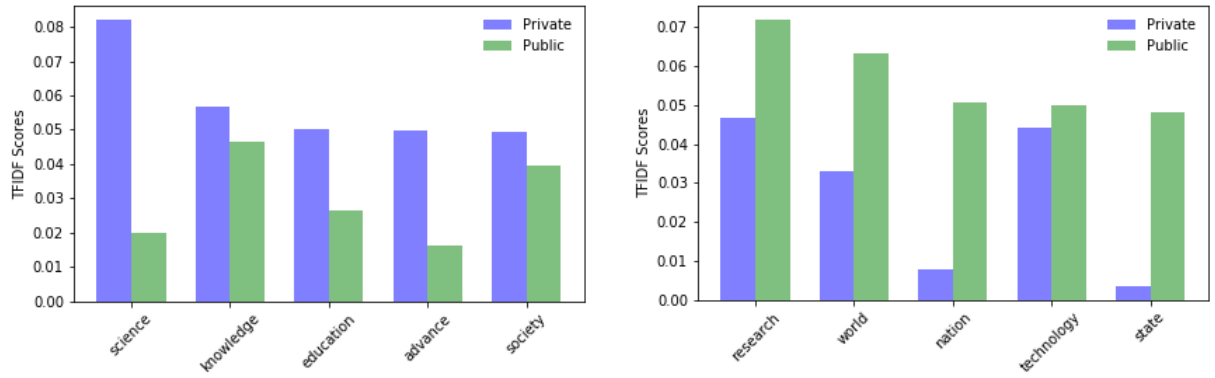
characterize relevance in larger corpora. In the Table 3 below the tf-idf scores for first 25 tokens with highest scores in both the classes (ie. Public and Private) are presented.

Table 3: Tf-idf scores for first 25 tokens with highest scores in both the classes (ie. Public and Private)

Private		Public	
science	0.081943268	research	0.071719091
knowledge	0.056594438	world	0.063129852
education	0.050094836	nation	0.05076635
advance	0.049776156	technology	0.049906213
society	0.049410098	state	0.04825906
research	0.046799363	knowledge	0.04637715
leader	0.045582243	develop	0.042591231
global	0.044523996	engineer	0.040120212
technology	0.044225111	society	0.039561888
educate	0.042067849	high	0.038210151
create	0.040750866	provide	0.036709806
applied	0.040355477	serve	0.034957449
prepare	0.039880354	quality	0.033608424
environment	0.039560895	benefit	0.03351393
provide	0.038891255	conduct	0.032934404
important	0.03871905	educate	0.032474353
community	0.033770225	leader	0.03237334
world	0.032976493	solution	0.030553474
problem	0.032368567	discovery	0.03022093
scholarship	0.030290559	graduate	0.028531669
leadership	0.028049867	innovation	0.028128905
learning	0.02741647	global	0.027577932
interdisciplinary	0.027188976	leadership	0.027501058
quality	0.027041394	service	0.027494654
address	0.026898937	education	0.026526552

Figure 1 shows a bar diagram plotting the tf-idf scores of the top 5 tokens of Private (Fig. 1-a) and Public (Fig. 1-b) schools. To facilitate comparison, the corresponding tf-idf score of the same token for the other class is plotted. For example, ‘science’ has the highest tf-idf score for Private schools. In Fig. 1-a the tf-idf score of ‘science’ for Public schools is also shown.

Figure 1: Top 5 (absolute tf-idf scores) tokens for Private and Public Schools.

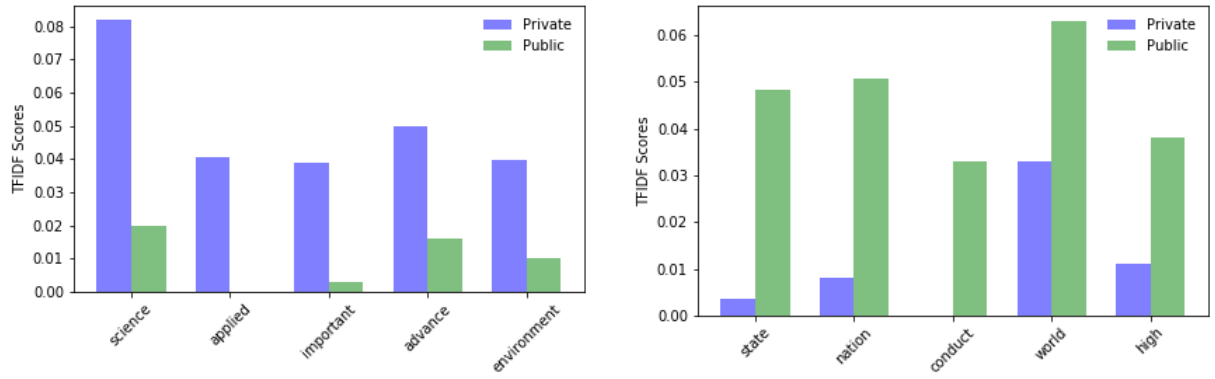


(a) Private **(b) Public**

Intuitively, these results make sense because public engineering schools which are predominantly part of State Universities were found to use words like state, nation, and world. In contrast, private engineering colleges were found to use words like science, education, and advance more than their public counterparts. Note that in Fig. 1-a tf-idf score for 'society' is almost equal. Similarly, for Fig. 1-b 'technology' for both the classes have similar score. This implies that these tokens occur with similar frequency and thus is not a good candidate for a feature which can discriminate between the two classes.

In the second iteration of assigning the tf-idf scores, we were interested in going beyond the raw tf-idf scores for tokens to calculate the differential tf-idf. This score was insightful in generating a list of most discernible words from either set of mission statements. Figure 2 shows the top 5 tokens for which the difference in tf-idf scores is maximum.

Figure 2: Top 5 (differential tf-idf scores) tokens for Private and Public Schools.



(a) Private **(b) Public**

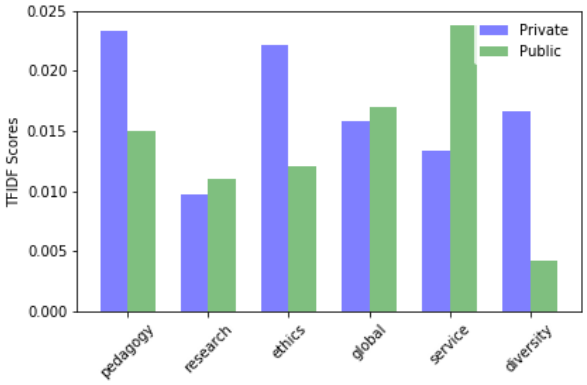
Figure 2-a shows that tokens like 'applied' which has a tf-idf score of approximately 0.04 for private schools has a score of 0 for public schools. This indicates that this token is a good candidate for discriminative feature for both the classes. Similarly, Figure 2-b shows that the token 'state' has a very high tf-idf score for public schools compared to private schools.

Discriminative features are useful for conducting machine learning based predictive analytics. For example, a future work of this research could be developing an automated classification system which can distinguish whether an input mission statement is of a public or private college.

2. Categories of Interest for Word Occurrences

There were six categories for which tokens of interest were generated from the dataset for further analysis. These categories, as explained earlier were based on prior work by Creamer and Ghoston (2013) in analyzing mission and vision statements of engineering colleges. As can be seen from the graph presented below, the scores for "Research" and "Global" seem to be comparable in terms of tf-idf scores assigned to the occurrence of tokens under these two categories. However, tokens grouped within "Ethics" found a higher tf-idf score in Private colleges as compared to public colleges. Similarly, "Diversity" was more prominent in terms of a higher tf-idf score in use in mission statements of Private colleges as compared to those for Public colleges. Although Private colleges likely emphasized on terms related to Diversity and Pedagogy and Ethics, Public colleges were higher in the category of tokens grouped as "Service". Figure 3 compares the tf-idf score for each of the categories for the two classes.

Figure 3: tf-idf Scores of each Category for Private and Public Schools.



Potential Threats to Validity

In this analysis, potential threats to validity include (a) threats due to complexities of semantics, and (b) threats due to number of uncontrolled variables.

(a)Threats due to complexity of semantics. As can be seen through the synonyms generated by WordNet in Table 2, some of the words generated as synonyms are not appropriate given the context of use for comparing public versus private institutions. For example, the Global category generated words such as world-wide, world, human-kind, etc. which may be relevant to the mission of a university, however, ball-shaped, globose, or globular which also were generated as synonyms were not of particular use in this context. This problem can be seen to be due to the complexities involved with semantics, in NLP systems. Jurafsky and Martin (2007) describe semantics as the knowledge of meaning. They note that language processing systems differ from other data processing systems since they use knowledge of the language. Training a machine to

understand knowledge and meaning is a non-trivial task. Jurafsky and Martin describe the complexities in lexical semantics (ie. the meaning of all words) as well as compositional semantics (ie. what words mean in relation to other words in a context). They provide an example of a sentence:

The blind bat swiftly flew past in the dark night.

He was shocked, but did not bat an eyelid.

In the first instance, the word *bat* refers to the flying mammal, while in the second instance it refers to a verb (that related with batting an eyelid). Thus, the meaning of the word *bat* is different for both the contexts depending on the use. In addition to challenges with semantics, the complexities of dealing with ambiguity is another major challenge for NLP systems, and systems are critiqued for having limited capabilities to resolve ambiguity. Ambiguity occurs due to multiple meanings for the same word. For example, the word *like* could be interpreted as a verb or an adverb, depending on the context it is used in. Jurafsky and Martin (2007) identify resolution of part of speech and word sense disambiguation as two important kinds of lexical disambiguation. As humans, we are often aware of the semantics and hence rarely have trouble with ambiguity, for a machine however, resolving ambiguity is an important concern, and is often difficult to train due to limited availability of context specific data. Jurafsky and Martin elaborate on the myriad challenges in the development of intelligent language processing systems by noting that:

> "language technology relies on formal models, or representations, of knowledge of language at the levels of phonology and phonetics, morphology, syntax, semantics, pragmatics and discourse." (pg.15)

Thus, interpretations drawn from studies involving automated systems must be done with caution. In this study, we used WordNet generated synonyms which are relevant for global terms, but may be lacking for engineering education contexts. Even after synonyms were generated, the synonyms are not exploited by the tf-idf method since tf-idf considers each word separately. One of the potential future works in this regard could thus be development of specific annotated corpora for engineering education contexts, to aid NLP research in this field, and use of modified tf-idf methods to cater to semantic similarities across words in the corpora.

(b)Threats due to number of uncontrolled variables. This study presents a descriptive overview of the words used by colleges in their mission statements. In order to achieve this descriptive statistical summarization, this research uses NLP methods. However, this research does not control for variables such as school size, geographical location, school purpose, etc. These additional variables may add another layer of complexity and interest to this analysis, and is proposed as future work.

Discussions and Implications for Future Work

This research presents an exploratory textual analytical approach to analyzing mission statements of engineering colleges. Our analysis shows that differences exist in the mission statements developed by private colleges versus public colleges of engineering. The results of this analysis

confirm prior analysis by Creamer and Ghoston (2013) who inductively reviewed mission statements for colleges of engineering. Similar to Creamer and Ghoston (2013) the results of this analysis revealed that colleges are committed to Service, Growth of Knowledge in areas such as Science, and Advancement in Technological Innovation.

Prior literature suggests none (Dale & Krueger, 2002) to little (Pascarella & Terenzini, 2005) difference in pay-off or occupational status, between students attending a more selective college versus a public. However, based on the analysis in this research, and if mission statements are truly representative of the underlying principles and intents of an institute, there does exist a reasonable difference in the priorities to various topics or values. For example, it seems from the results of this exploratory analysis that the public institutions valued terms synonymous with "service" such as outreach, and emphasized in their mission statements words like "state", "nation", and "conduct". Similarly, analysis of the corpora for the private colleges found more instances of terms and phrases synonymous with the categories of "pedagogy", "ethics" and "diversity", while using of words like "science", "advancement", and "environment". These highlighted differences in values between the two types of universities may indeed have an impact on the student experience through attending a four-year engineering degree program.

This research has strong implications for the future of engineering education research by successfully demonstrating how NLP can be used for summarizing data from large textual corpora. This strategy of employing NLP techniques for quick descriptive overviews of large textual datasets can be of interest to administration for gaining faster overviews related to their policies based on available textual data. This research can also be useful for engineering education researchers who are interested in analyzing large corpora with constraints of limited time and resources.

Future work in this area can help to gain a deeper understanding of comparisons across public and private engineering colleges by including more mission statements from colleges as part of the analysis. As described in an earlier section on threats to validity, another area of future work is in terms of developing engineering education specific datasets which are more relevant to contexts of engineering, and more relevant to the engineering community of practice.

References

Barnett, R. (2003). *Beyond All Reason: Living with Ideology in the University*: ERIC.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*: " O'Reilly Media, Inc.".

Church, K. W., & Mercer, R. L. (1993). Introduction to the special issue on computational linguistics using large corpora. *Computational Linguistics, 19*(1), 1-24.

Creamer, E. G., & Ghoston, M. (2013). Using a mixed methods content analysis to analyze mission statements from colleges of engineering. *Journal of Mixed Methods Research, 7*(2), 110-120.

Dale, S. B., & Krueger, A. B. (2002). Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables. *The Quarterly Journal of Economics, 117*(4), 1491-1527. doi:10.1162/003355302320935089

Delucchi, M. (1997). "Liberal arts" colleges and the myth of uniqueness. *The Journal of Higher Education, 68*(4), 414-426.

Detomasi, D. (1995). Mission Statements: One More Time. *Planning for Higher Education, 24*(1), 31-35.

Fellbaum, C. (1998). WordNet: an eletronic lexical database. Cambridge, Massachusetts, EUA: The Mit Press.

Fox, N., Ehmoda, O., & Charniak, E. (2012). Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate. *Proceedings of the Georgetown University Roundtable on Language and Linguistics (GURT)*, 363-371.

Jurafsky, D., & Martin, J. H. (2007). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (Vol. 26).

Kibuuka, H. E. (2001). Vision and mission statement in Christian higher educational management in Eastern Africa. *Journal of Research on Christian Education, 10*(1), 87-114.

Kreber, C., & Mhina, C. (2007). The values we prize: A comparative analysis of mission statements. *Higher Education Perspectives, 3*(1).

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods, 44*(3), 608-621.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41.

Morphew, C. C., & Hartley, M. (2006). Mission statements: A thematic analysis of rhetoric across institutional type. *The Journal of Higher Education, 77*(3), 456-471.

Newsom, W., & Hayes, C. (1991). Are Mission Statements Worthwhile? *Planning for Higher Education, 19*(2), 28-30.

Pascarella, E., & Terenzini, P. (2005). How college affects students (Vol. 2). KA Feldman: San Francisco, CA: Jossey-Bass.

Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide Note Classification Using Natural Language Processing: A Content Analysis. *Biomedical Informatics Insights, 2010*(3), 19-28. doi:10.4137/BII.S4706

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016). *Forecasting student achievement in MOOCs with natural language processing.* Paper presented at the Proceedings of the Sixth International Conference on Learning Analytics & Knowledge.

Savoy, J. (2015). Text clustering: An application with the State of the Union addresses. *Journal of the Association for Information Science and Technology, 66*(8), 1645-1654.

Shi, C., Xu, C., & Yang, X.-j. (2009). Study of TFIDF algorithm. *Journal of Computer Applications, 29*(6), 167-180.

Tierney, W. G. (1999). *Building the responsive campus: Creating high performance colleges and universities*: Sage.

Variawa, C. (2014). *Investigating the Language of Engineering Education.* University of Toronto.

Variawa, C., McCahan, S., & Chignell, M. (2013). *An Automated Approach for Finding Course-specific Vocabulary.* Paper presented at the 2013 ASEE Annual Conference & Exposition, Atlanta, Georgia.

Young, R. B. (2001). Colleges on the cross roads: A study of the mission statements of Catholic colleges and universities. *Current Issues in Catholic Higher Education, 21*(2), 65-81.

Appendix A:

List of Universities from where the Mission Statements for Engineering Schools/Colleges were Extracted

Name of Parent University (Private)	Name of Parent University (Public)
Stanford University	Georgia Institute of Technology
Massachusetts Institute of Technology	University of California Berkeley
Northwestern University	University of Michigan
John Hopkins University	University of Illinois at Urbana Champaign
Columbia University	University of Texas at Austin
University of Southern California	University of California, Los Angeles
Princeton University	University of Wisconsin, Madison
University of Pennsylvania	Texas A&M University
Cornell University	University of Virginia
Carnegie Mellon	University of California, Davis
Vanderbilt	Purdue University
California Institute of Technology	University of Colorado, Boulder
Yale University	University of Washington
Brown University	University of Maryland, College Park
Rennselaer Polytechnic Institute	University of California, Irvine
Duke University	University of California, Santa Barbara
University of Rochester	North Carolina State University
Case Western University	Iowa State University
Boston University	University of Pittsburg
Tufts University	University of Iowa
University of Notre Dame	Virginia Polytechnic and State University
Washington University of St. Louis	University of California, San Diego
Harvard University	University of Arizona
Lehigh University	University of Florida
Syracruse University	Rutgers, The State University of New Jersey
Tulane University	State University of New York at Buffalo
Rose Hulman University	University of Massachusetts Dartmouth
University of Miami	University of Tennessee, Knoxville
Santa Clara University	University of Minnesota, Twin Cities
	University of Connecticut