Asta Habtemichael

BIO 539

Final Project Paper

**Title**: Text Mining on Students' Reflective Essays

**Introduction**

Negatively stereotyped students, particularly students of marginalized groups, in an academic setting find it threating (G.L. Cohen and Garcia, 2005; Steele, 1997). Minority students generally have to deal with negative stereotypes about the intellectual ability of their group. Different studies had shown that this "identity threat" may cause stress (Blascovich et. al., 2001)due to the lack of sense of belonging and can also directly impact their academic performance Steele and Aronson, 1995). The impacts of identity threat are observed in different ways such as interpersonal negative relational consequences as exhibited through anxiety, social disconnection, and feelings of alienation (Cook et. al., 2012).

Self-affirmation exercises had been shown to be one effective way of buffering students against such stressors, even were connected with better academic performance (Shnabel et. al., 2012). Self-affirmation refers to major events of greater worth in a person's life that might be related to behavioral or cognitive roots which bolden self-worth (Sherman and Cohen, 2006). These values act as shield for the self and help individuals to reduce psychological threat and stress as well as improve their academic performance (Martens et. al., 2006; Cohen et. al., 2006).

This project will investigate the word choices of students of color in their reflective essays titled 'I believe' and how that could be used to evaluate the values they reflect upon. Word choices could be used as indication for understanding the establishment of sense of belonging as well as world constural of individuals based on their social and physical interaction with their environments.

**Hypothesis**

In this study it is hypothesized that the word selection of students reflects their core values to which they give great importance in their world construal. The hypothesis is based on the assumptions that;

\**All other factors are insignificant except the reflective exercise and core values of the students*

*\*Metacognitive and motivational aspects are fused for the purpose of building the principal core values*

*\*Events that had happened in their earlier life stages take precedence on setting values affirmed.*

The aim of this study is to highlight how inclusive pedagogy can be designed based on individualized curriculum which are built around the core values affirmed by students. The project is significant as it may add upon the science of pedagogy and curriculum design by filling the gap by understanding the connection between the values affirmed and the word choice that might help early identification of students' sense of belonging in an academic setting. Furthermore, knowing the words related to such values enhances the ability for academic advisors to engage the students and stimulate their attention by focusing on issues that are more salient and intriguing to capture students' interest.

**Methodology**

An exploratory analysis of the essays was performed on the text body of the data. Below are the steps of the methods employed from sample collection to analysis.

**Sampling**

The sample essays were collected from freshman students in an introductory biology course where 85% of the students were students of underrepresented groups in STEM fields. The essays were written in their first week of classes in an open-ended prompt titled "I believe." They were asked to write a reflective essay of minimum 500 words about major events in their life that had some kind of attribute towards their choice to be in a STEM field or their future career that they are planning to pursue.

There were 153 students in this particular section for that particular term and all students had submitted an essay that was considered for further analysis for the purpose of this project.

**Data Analysis**

The data analysis part where the work on the data was performed using Pandas in Phyton is discussed below.

**Data Loading**

The dataset contains id tag for each essay or text document at the beginning of the essays. The data was not preprocessed; therefore, all the essays came as separate documents. An R multiple filed reader 'readtext' package was used to assemble the dataset into a single .CVS file where the original texts were contained in a column.

The data as .CVS file was read into Pandas in using Phyton for further text analysis.

**Cleaning**

As with most datasets the data was not ready for processing as they were, had unwanted markings and text features that needed to be cleaned. First step was cleaning the id tag from the text body, in order to do that we set the data type for the data frame as string and we remove the first 11 digits. Then pandas *clean* function for string was arranged to reconstruct the line ending, and removal of text mining impairing features in the text like URLs, Markdown, underscores, and multiple quotes.

The clean data was directed to be stored in a new column 'selftext\_clean' respective of the rows the content was in the raw data through a for loop that iterates the data set and applies the cleaning. Once the data was cleaned it was ready for further text analysis using linguistics operation.

**Natural Language Processing (NLP)**

The 'spaCy' language package for English language was used-of-speech (POS) tagging and finding different kinds of words. For this project nouns, verbs, and adjectives were identified from the text and were saved in columns in the data frame based on the coding reported in GitHub post by Albrecht et. al., (2019).

In this project Exploratory Data Analysis (EDA) for text data was employed to systematically go through the data and get some intuition on the distribution and hidden correlations of the data. Toolsets including summary statistics and frequency counts for categorical features were utilized to analysis the data. When dealing with text data, the main focus of the analysis work is on frequencies. Individual words were tokenized in order count the word usage and find its frequency. The individual values of word frequencies were visually presented using word cloud for a pictorial depiction and bar chart for numerical representation.

**Results and Discussion**

The basic data exploratory analysis done after the clean up step assured that the raw data and the clean data are consistent in terms of data structure, data type, and data frame size. The statistical summaries computed using pandas *describe* method for the columns of the data frame of categorical and numerical nature are as follows;

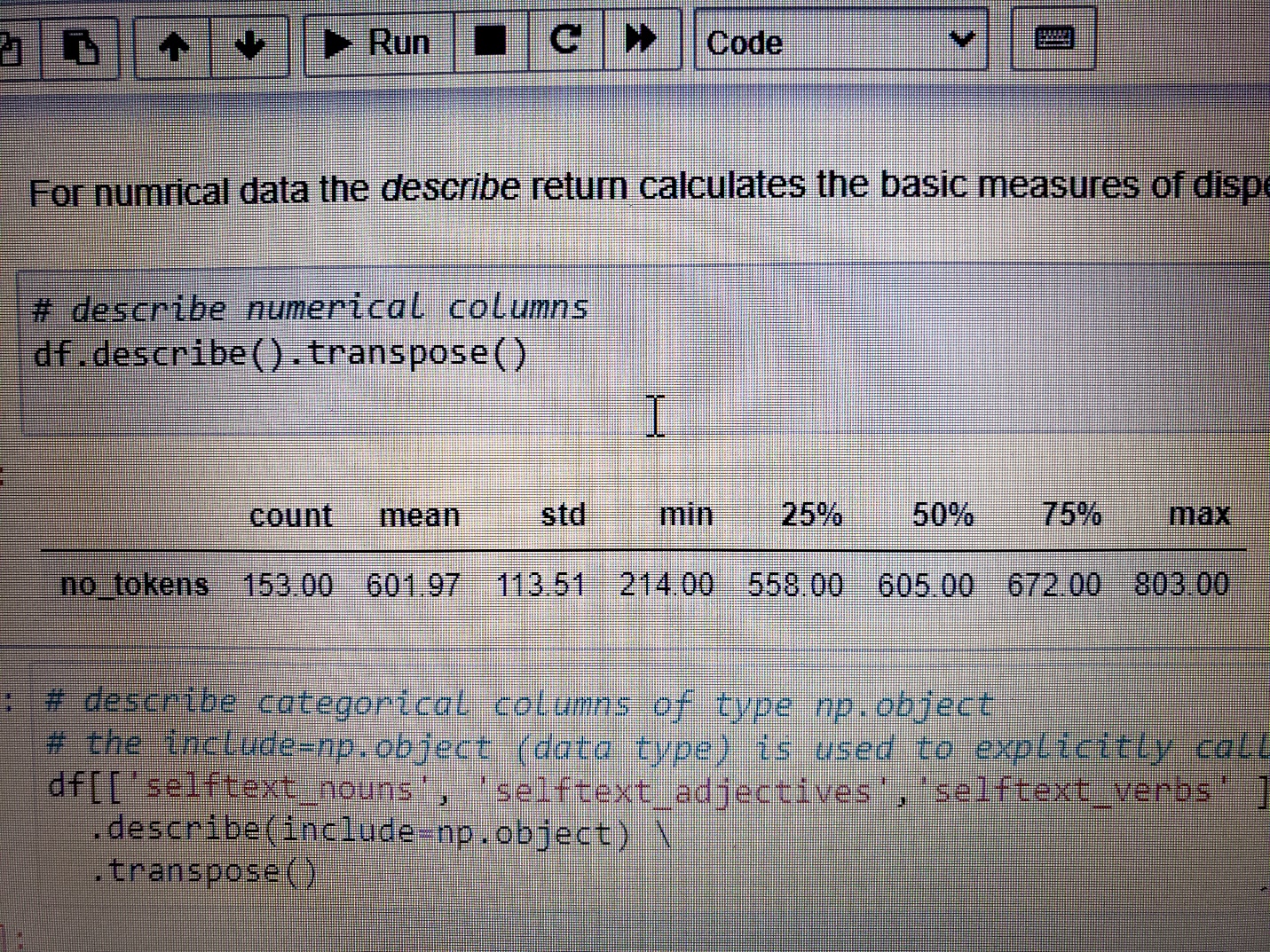


Fig.1 The Five-number summary and measures of dispersion for the numerical value *number of tokens* in the data.

The token value in the statistical summary represents number of words used by the students and in the count column it represents the number of rows in the data set. The returned result shows that the mean values of tokens (words) to be 601.97 in the data set where a maximum of 803 tokens (words) are recorded(Fig. 1.). That is the students used an average of 601 words in their essays.

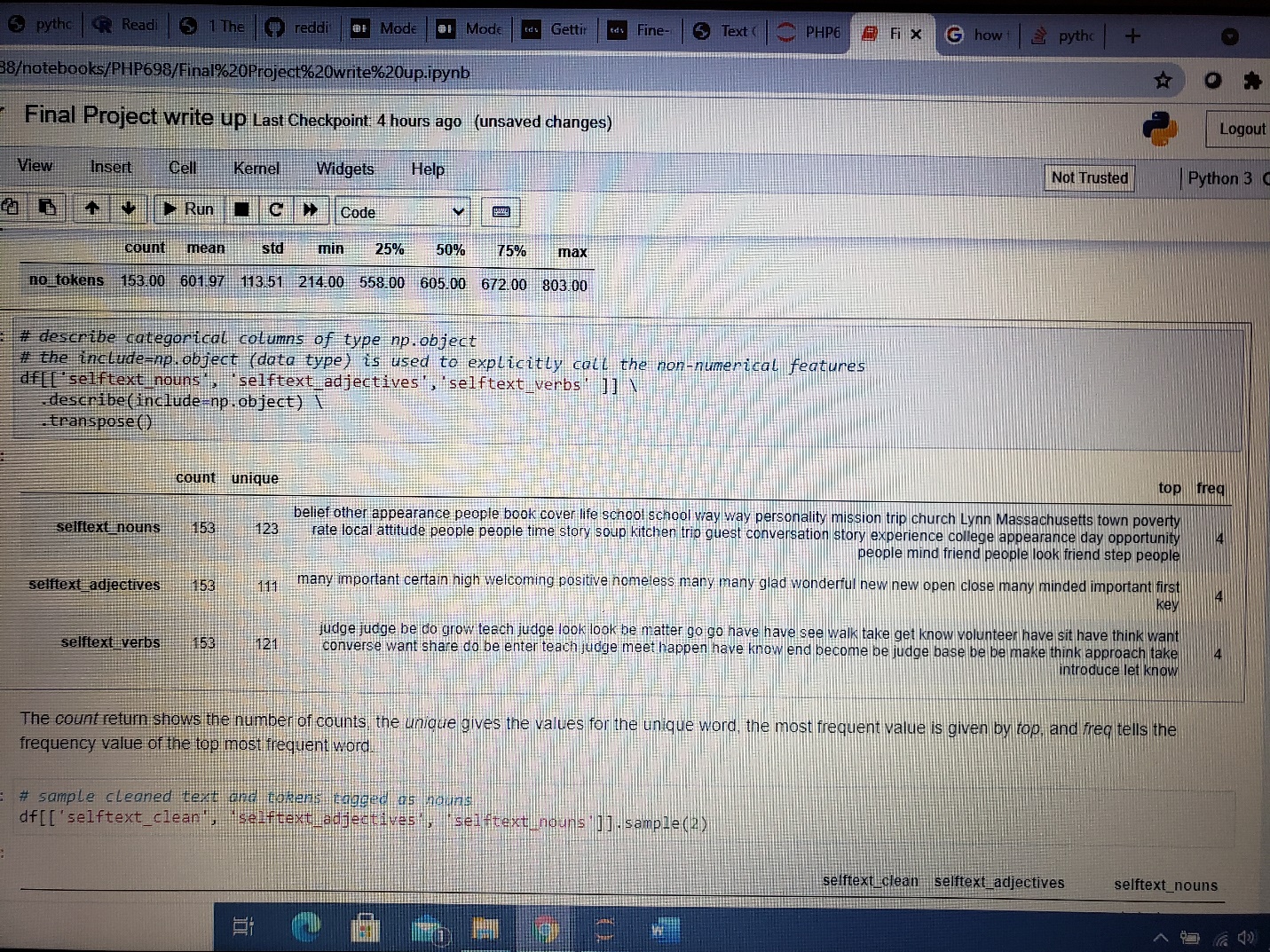


Fig.2. The summaries for a categoriacal columns in the data set explicitly called for specific columns of selftext\_nouns, selftext\_adjectives, and selftext\_verbs.

The summary metrics or categorical data are different from numerical data in the pandas *describe* method. The *count* values show the count of the rows for that column, *unique* shows the values for the unique tokens in that column, while *top* gives the most frequent tokens in that column, and *freq* indicates the frequency of the most frequent value (Fig. 2.).

In this project were more concerned about the word usage and what the frequency of the words looked like. In order to accomplish this task, the words in the essays were tokenized (made into a list) using a *my\_tokenizer* function which employs *split()* to individually set words as a token based on the word spacing in the essays. It shall be noted that the resulting list needed to be cleaned from ‘stop words’, basic linguistic components like conjunctions, prepositions, articles, and common words which are less significant as the removal of these words will not affect the significance of the message in the essay.

The word (token) frequency analysis was limited to adjectives and nouns as these are the words that correlate with our research question which focuses on identifying values affirmed and feelings of belonging. Nouns could be related as to what value is important for our sample students by looking at the most frequently used nouns in terms of the essay attributes namely impactful event or career goals.

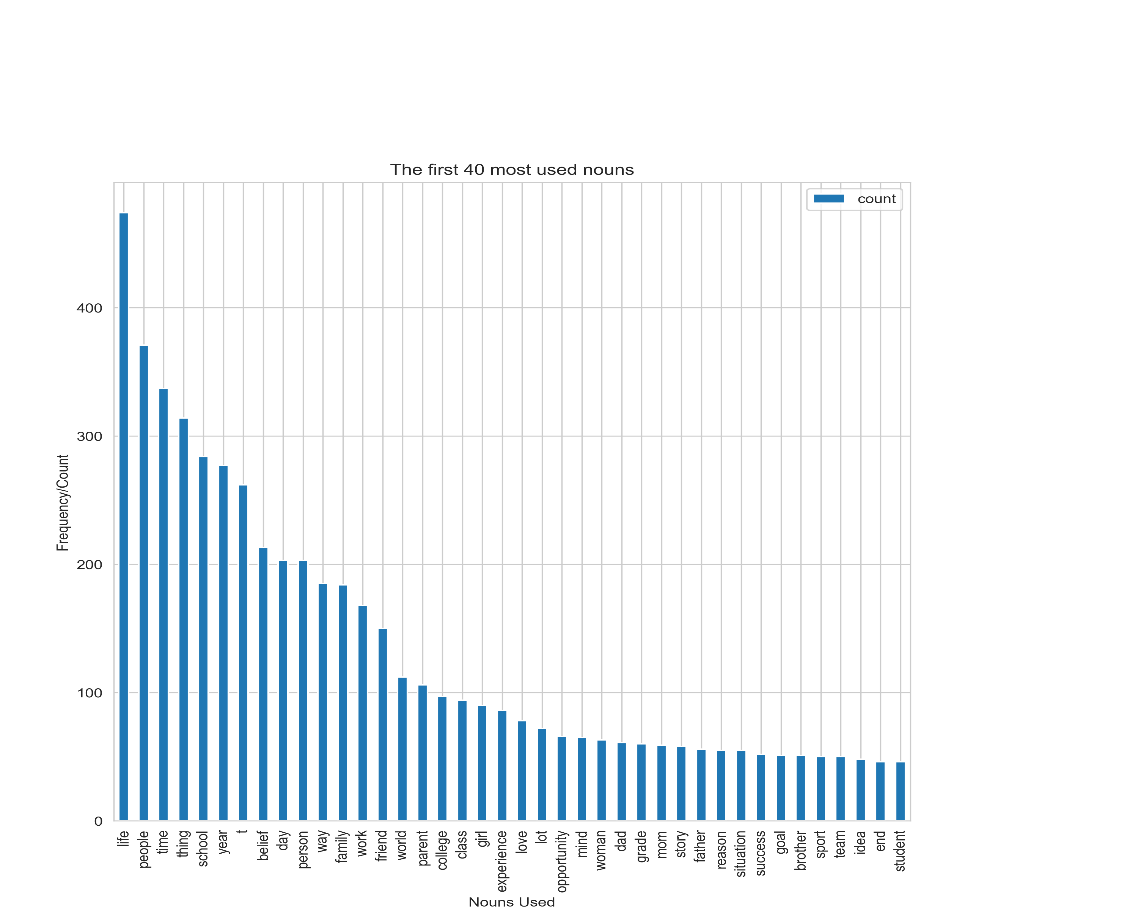


Fig. 3. The count/frequency of nouns used by students in their essays. The first forty nouns are plotted against their respective counts.

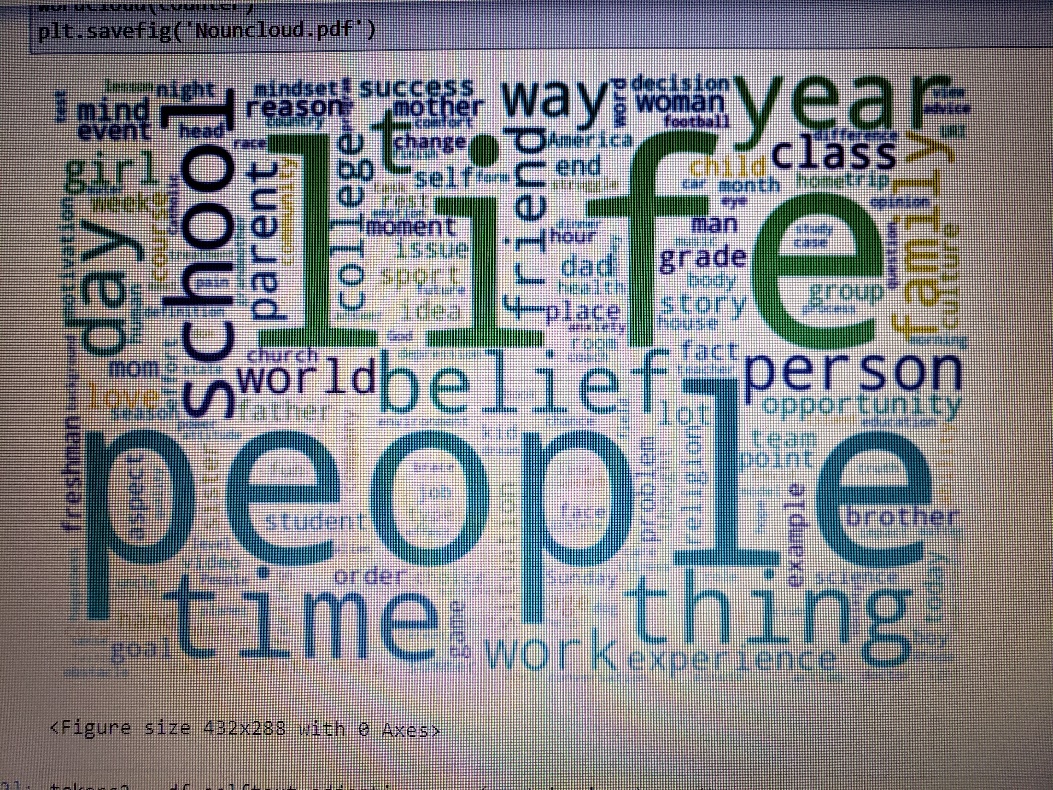


Fig.4. A world cloud visualization of the highly frequented nouns in the essays.

As we can see in fig.3 and fig.4 the first two nouns that score the highest count are life and people. A list of words such as time, year, thing, school, belief, person and friend making part of the ones that are jumping out as the most common ones in the essays. Basically, we can understand that the two attributes of the essay prompt, impactful events and career goals, are largely connected with people, life, things around the participant students. This indicates that extrinsic factors matter the most in guiding a student’s career choice and significantly affect how they relate to their current academic path as well. The cores of grade, sport, student, even success being in the lower ranges (See fig.3 and fig.4) further strengthens our analysis that their values are more connected to extrinsic factors that may be influenced by metacognition and motivational factors rooted in their social and cultural contexts.

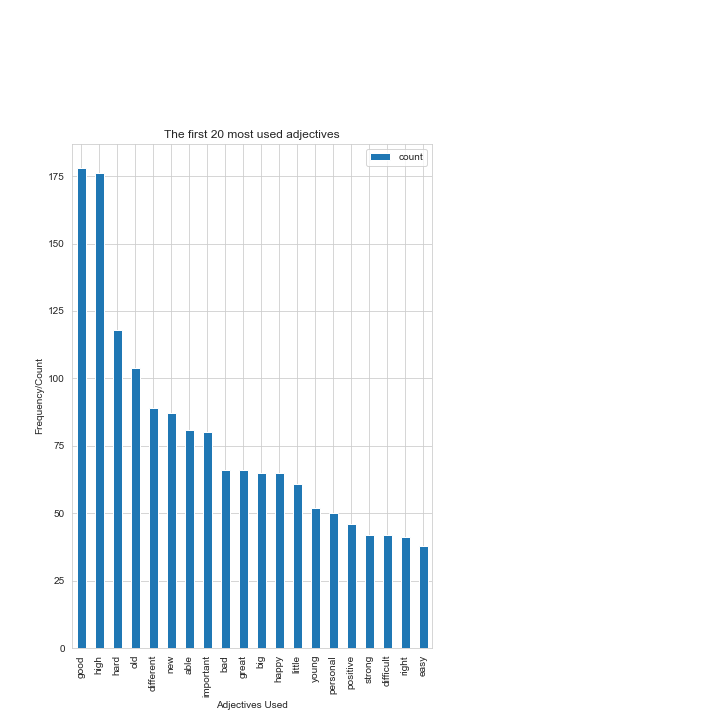


Fig.5. The most frequently used twenty adjectives in the essay prompts of this project against their count values.



Fig.6. Word Cloud visualization of the most frequently used adjectives in the essay prompt exercise.

The adjectives with the highest frequency value high and good (176 and 175 respectively), followed by hard, old, different, and new can be related to the students emotional and psychological interactions to the impactful events that shaped their career goals or how their career goal is reflected in their future (Fig. 5 and Fig.6). Further analysis is needed to connect the individual sentiment of the students towards their future goals as connected to the major event in their life to fully understand the value affirmation attribute connect with that particular event and how that translates into a coping mechanism or shielding the self from identity threats.

However, based on the connections we have based on the word choice and usage frequency we can generally attribute to what values and feeling are to guide their career path choice. The adjectives which are related to challenges and tough times, such as hard, old and different are comparable in range of frequency with words that indicate good tides of change such new, able and great ultimately indicating a change of pattern in their life (See Fig.5 and Fig.6). These positive attributes of change when connected with the values that they affirmed identified via the nouns frequency, it might be used as an overarching indication of students sense of belonging in the current environment, which is academia at the time the essay was written.

**Conclusion**

This report intended to introduce basic text data exploration and text analysis with a goal of identifying the word choice and word frequency used in essay prompts. Techniques useful for data manipulation, transformation, and visualization were used to identify words of significance in relation to the research question. Words were tokenized and categorized into nouns, adjectives, and verbs in order to relate them to their respective frequencies and enable us to identify their significance in the essay as attributed to our research question. Words of higher usage frequency were identified from both adjectives and nouns that were used to link to values affirmed and the significance of these values in career path choice. Further research is needed to fully understand the significance of the affirmed values in relation to the career path choice, but preliminary analysis in this data set indicates social and cultural context of students (reflected by words life and people) and paly a great role as part of guiding their career path choice in relation to impactful events in their life. This report also opens the door to exploring more attributes of sense of belonging via the adjective analysis (representing feeling) utilizing sentimental values as related to the values affirmed. The way forward would consider building a model that accounts for both intrinsic and extrinsic factors in determining student’s choice of career path.

**References**

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. (2001). African Americans and high blood pressure: The role of stereotype threat. Psychological Science, 12, 225-229.

Cohen, G. L., & Garcia, J. (2008). Identity, belonging, and achievement: A model, interventions, implications. Current Directions in Psychological Science, 17, 365-369.

Cohen, G. L., Garcia, J., Apfel, N., & Master, A. (2006). Reducing the racial achievement gap: A social-psychological intervention. Science, 313, 1307-1310.

Cook, J. E., Purdie-Vaughns, V., Garcia, J., & Cohen, G. L. (2012). Chronic threat and contingent belonging: Protective benefits of values affirmation on identity development. Journal of Personality and Social Psychology, 102, 479-496.

Martens, A., Johns, M., Greenberg, J., & Schimel, J. (2006). Combating stereotype threat: The effect of self-affirmation on women’s intellectual performance. Journal of Experimental Social Psychology, 42, 236-243.

Sherman, D. K., & Cohen, G. L. (2006). The psychology of selfdefense: Self-affirmation theory. In M. P. Zanna (Ed.), Advances in experimental social psychology (Vol. 38, pp. 183-242). San Diego, CA: Elsevier Academic Press.

Shnabel, N., Purdie-Vaughns, V., Cook, J. E., Garcia, J., & Cohen, G. L. (2013). Demystifying Values-Affirmation Interventions: Writing About Social Belonging Is a Key to Buffering Against Identity Threat. Personality and Social Psychology Bulletin, 39(5), 663–676.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. American Psychologist, 52, 613-629.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. Journal of Personality and Social Psychology, 69, 797-811.

Most codes used for this project were adopted from <https://github.com/datanizing/reddit-selfposts-blog/blob/master/02-data_exploration.ipynb> and were amended as necessary.