Analysis of an Extracted Discipline-Specific Computer Science Vocabulary List

August 13, 2021

Adya Dutt, Carrie Demmans Epp

Abstract

A strong vocabulary allows for more understanding. Vocabulary can be categorized into 3 tiers. Tier 1 contains commonplace and basic vocabulary. Tier 2 words are used by mature speakers across a variety of subject matters. Tier 3 contains low-frequency words that are used in specific disciplines. To gain conceptual knowledge in academic domains, a strong foundation of Tier 3 vocabulary is essential. Many disciplines do not have a vocabulary list in place, making the learning process more difficult. Computer science is one such discipline, which is why this project aims to create a method for automatically creating a computer science vocabulary list. Starting with a premade computer-science corpus, a computer-science vocabulary list is extracted using computational linguistics techniques. This includes removing unrelated words from the list by extracting text from a compilation of non-computer science resources and using the extracted text to remove any similar words from the original vocabulary list. The resulting list could help

enhance student learning in computer science and this process could be duplicated for creating another discipline-specific vocabulary list.

Introduction

Comprehension is necessary to learning new content. When reading, words are the foundation to comprehending a text. New content can be understood easily if a robust vocabulary is present. For content that is specific to a discipline, the reader needs to know the specialized vocabulary of that discipline to understand its texts. Therefore, strong academic vocabulary is necessary to excel in any area of study, as comprehension correlates to proficiency, especially at the university level. English vocabulary can be broken down into 3 tiers that describe how widely the words are used. These tiers help explain understanding. Tier 1 are common, general-purpose words that are used daily (e.g. school, eat, fan). Tier 2 vocabulary are higher-level words used across many domains (e.g. cite, parallel, analyze) and by mature English speakers. Tier 3 vocabulary are discipline-specific words and are specialized to only a few contexts (e.g. boolean, isotope, photosynthesis). This tier breakdown has been used by researchers to create various wordlists for all tiers (Bhatia, 2018). However, Tier 3 wordlists cannot be compiled as easily as the words they contain are used infrequently in comparison to Tier 1 and 2.

Disciplines like computer science have no such vocabulary list. The goal of this project is to use prior research and natural language processing techniques to create a wordlist of computer-science vocabulary and distinguish it from non-computer science vocabulary. This computer-science vocabulary list can be used to improve student comprehension. It can also be applied to teaching curricula and could be used to analyze students' computer-science word usage in online forum platforms.

Methodology

This project adapted the previous work and progress reported in “The Computer Science Word List - Extraction and Analysis” (Bhatia, 2018). A corpus comprised of publications from all different subareas of computer science was already constructed. This corpus was used to extract all the words from the corpus.

This was done by compiling the needed code files, creating input, output, and removal folders to organize the directory, and downloading all twelve source texts which contained the corpus categorized by subject. Spreadsheets were also created and prepared before running the code to generate the output. Steps 1 to 3 were used for pre-processing, where the text was prepared to extract the computer science list. Steps 4 to 6 were for the analysis of the text; unrelated words were removed.

Pre-processing

To run the 6 Python code files, all needed libraries were downloaded. The first code file separated the blocks of text in each source file by organizing one word on every line and then saving it in a new file.

The second code file organized words from each source text by alphabetical order and a new file was created separately for each letter of the alphabet. The numbers of sources listed in each code file also had to be altered to include sources 3-12.

The third code file created a list of headwords for each source file from all the words present. Headwords are the derivative of related words with different suffixes. This is done because

learners can comprehend headwords with more ease and can make connections to related words (Bhatia, 2018). A generated dictionary of words and their headwords, created using the Range software, was used from the previous research to remove out of vocabulary (OOV) words and keep the words and their headwords. The order of the code in the if statement and the for loop had to be altered, as well as some outdated functions.

Analysis

The fourth code file compiles all the headwords into one file and counts the frequency of the headwords present in each source file. The result was a list that contained words from all tiers along with their frequencies and the range among all 12 sources.

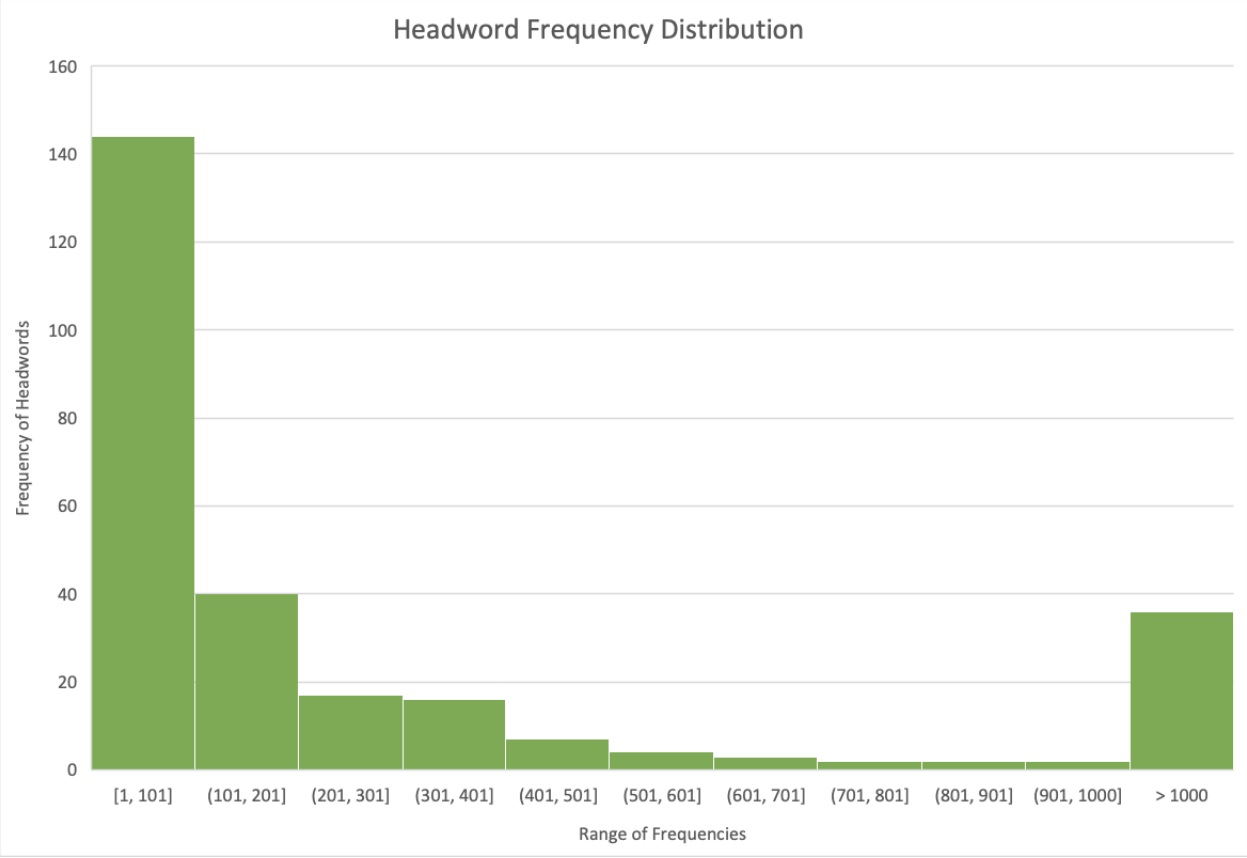

Figure 1.1: Histogram of headword frequency distribution.

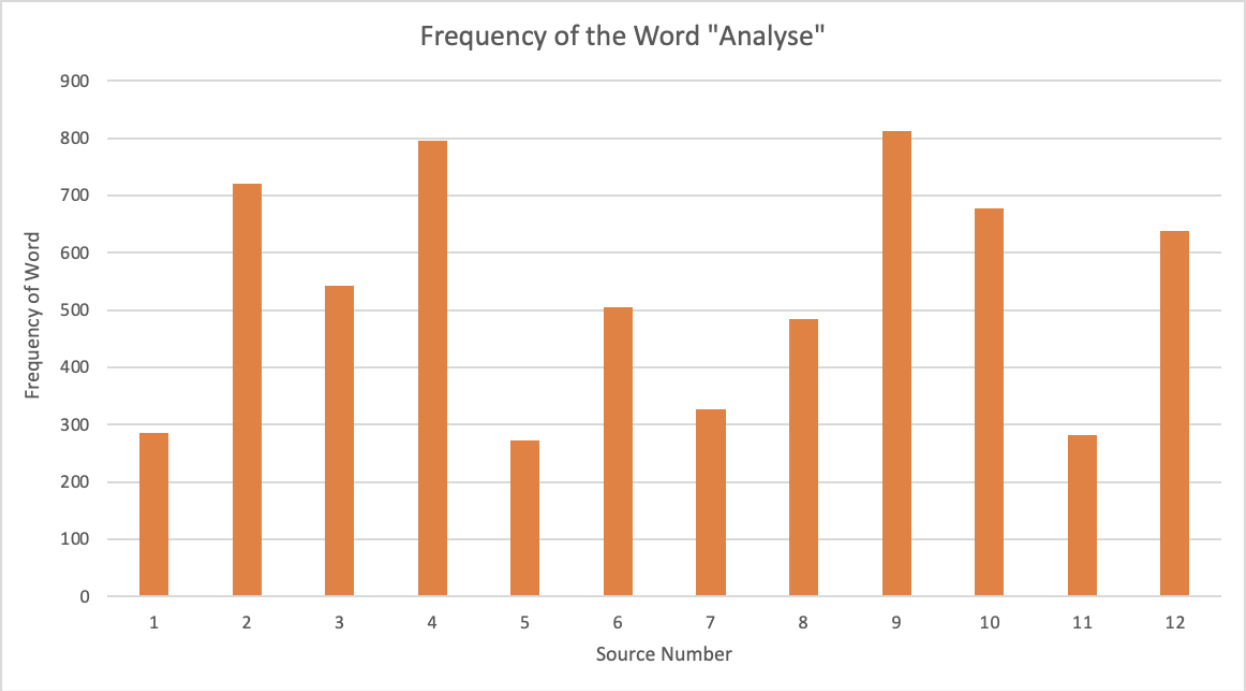

Figure 1.2: Frequency of the word "Analyse" across the source texts.

The goal of the next code file was to eliminate any lower-tier words. However, the fifth code file had numerous errors, and it could not run until a list of tier 1 and 2 words were compiled. The New General Service List (Browne et al., 2016) and the New Academic Wordlist (Coxhead, 2000) were composed of the most common words in the English language, meaning these lists were compiled as Tier 1 vocabulary. Tier 2 vocabulary lists were also compiled (*Tier 2 Vocabulary Lists*, 2018), where words were based on different subject matter areas such as science, math, art and different academic years (*Marzano Vocabulary Lists by Grade Level for LA, Sci, SS, and Math*, 2018). All these sources were either PDF and spreadsheet files. After downloading these files, I wrote additional code to read the PDFs and extract text from them. The library pdfplumber was used, as it had the ability to read tables as well, which were present in some of the PDF files (Singer-Vine, 2015). To extract the text from the Excel spreadsheets, only the words on one specific column and sheet were needed. I wrote code to read through

every line and drop all other values that were not necessary. Each file was read and all text was extracted into one new file, where one word was present on each line. All words were capitalized as well for consistent coverage that was readable to the computer.

The fifth code file was then run, and all tier 1 and 2 words that were collected in the last step were removed from the list of headwords. The sixth code file combined the final list of headwords and their frequencies, which was then imported onto a spreadsheet for the final computer science wordlist. The complete code was saved and committed in a Github repository.

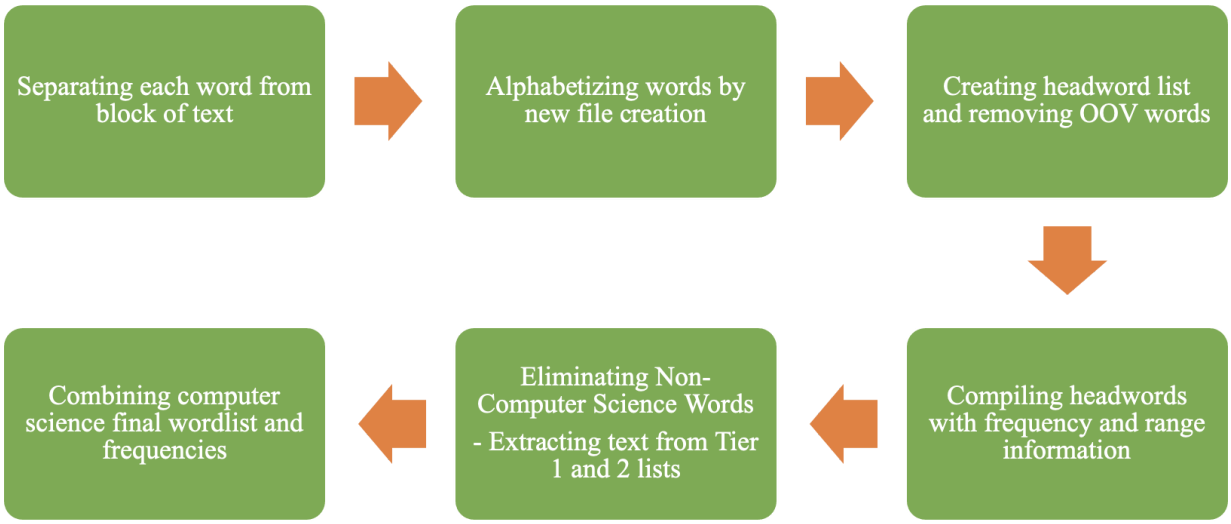

Figure 1.3: Flowchart of the execution process.

Results

The final wordlist contained errors, as some non-computer science words were present and some computer science words, originally in the corpus, were not present in the list. After analyzing the output, the reasoning behind the few missing computer science words seemed to arise from the generated dictionary of headwords, as any OOV words are removed, including computer science

words. These missing computer science words were mostly acronym-based (ex. SQL, CTRL, AI), and these were not included in the headwords, thus categorizing them as OOV. The non-computer science words in the final list were due to the Tier 1 and 2 lists not containing those words. Names of countries were the most frequent non-computer science words. Finding more Tier 1 and 2 lists, such as lists for country names, could be a viable solution to this issue. These lists could then be used to remove unrelated words from the final wordlist.

Headwords in Final Wordlist	OOV Words Missing From Final Wordlist	Non-Computer Science Words in Final Wordlist
273	209	219

Table 1: Analysis of final wordlist.

Discussion

As mentioned above, the generated dictionary of headwords and the tier lists had limited vocabulary and excluded certain words that could potentially be in the computer science wordlist. Combining the generated dictionary with the OOV words would allow all words to become headwords, while still maintaining the derivative grouping. Using more tiered vocabulary lists would decrease the amount of non-computer science words in the final list. Additionally, since there is no established method to create a discipline-specific wordlist, the wordlist is more flawed, with missing or incorrect tier words. Biases in tier structure also exist, as categorization of words is based on one's familiarity with the discipline. This is why manual

alterations had to be avoided as much as possible, making it more difficult to create a methodology for discipline-specific wordlists.

Future Work

This computer science vocabulary list can be used to help teach students computer science. It can be used to improve student understanding throughout a course by analyzing the context these words are used in. Online platforms like forums or a virtual classroom can use this wordlist in an automated way, where a machine could monitor the use of the vocabulary and track progress. Computer science textbooks or learning resources could also use these lists to categorize discipline-specific vocabulary. This wordlist methodology is not limited to computer science; any other discipline can also create its own wordlists.

Acknowledgements

I would like to thank my principal investigator, Carrie Demmans Epp, for taking the time to guide me through my research. She has been extremely supportive throughout my internship and was always open to any questions or ideas I would have. I sincerely appreciate the valuable insights I have received from her on topics varying from conducting research to studying abroad. I have become much more confident in my computer science and professional skills through her advice and encouragement.

I would also like to extend my gratitude to Daniela Teodorescu, who supervised my participation in the Cree Corpus side project. Through our check-in meetings, Daniela has been immensely helpful and has provided me with reassurance on countless occasions.

In addition to the appreciation I have for this amazing opportunity that was provided by WISEST, I want to thank the incredible Summer Research Program coordinators for all your hard work in hosting such a remarkable program and implementing a supportive environment. I have developed professionally, socially, and personally through the numerous PD and networking sessions.

A special thank you goes to the Motorola Solutions Foundation for sponsoring my time at the Summer Research Program. Their generous support has allowed me to participate in this profound program.

Lastly, I would like to recognize my fellow 2021 Summer Research Program peers. All my peers have helped establish a comprehensive environment this summer which I have cherished.

References

Bhatia, T. (December 2018). *The Computer Science Word List - Extraction and Analysis* [Unpublished thesis]. Goa Campus, Birla Institute Of Technology And Science Pilani.

Browne, C., Culligan, B., & Phillips, J. (2016, May). *New General Service List* (1.01 Alphabetized) [Dataset]. New General Service List Project. http://www.newgeneralservicelist.org/

Coxhead, A. (2000). *A New Academic Wordlist* (1.0 Lemmatized For Research) [Dataset]. Victoria University of Wellington. https://www.wgtn.ac.nz/lals/resources/academicwordlist

Marzano Vocabulary Lists by Grade level for LA, Sci, SS, and Math. (2018). Sealyisd. https://www.sealyisd.com/common/pages/DisplayFile.aspx?itemId=2339209

Singer-Vine, J. (2015). *pdfplumber*. Github. https://github.com/jsvine/pdfplumber

Tier 2 Vocabulary Lists. (2018). Hyde Park Schools. https://www.hpcsd.org/site/default.aspx?PageType=14&DomainID=27&PageID=16944&ModuleInstanceID=23854