Common words and their frequency in the Wikipedia summary of two words

Subeen Kim

Project Overview

Wikipedia provides an excellent description of a particular word. However, I have never seen a function like comparing or showing the relationship between two or more words. So I created a program, which searches the summary part of Wikipedia for a common source of two different words, can show the common words and its frequency. The program takes each summary as a string type and then divides it into several words in list type. In the meantime, the program refines the applicable words for post-processing. Finally, the list of the two words is compared and the common word and frequency are displayed as a histogram.

Implementation

The program is composed of three different functions. The first function, <text_to_word> aims to refine the word being available to do further process. The function replaces non-alphabet characters into space or blank as well as change all into lower character. The string typed text is divided into multiple words by using split function. The second function <histogram> produces a dictionary composed of common word from two summary in which the key and value is the word and the number of word in the summary.

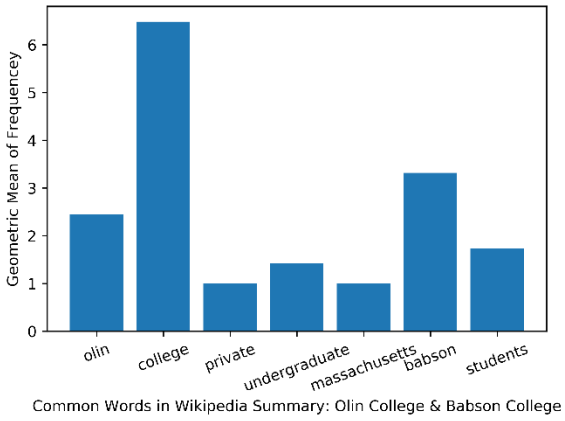
The third function, <matching summary>, is the main function of this program. The function gets two arguments, which will be each summary of two words. The arguments are refined by <text_to_word> function, then it is used to construct the histogram of each word through <histogram> function. For the common words which are both in each summary only can be stored inside of common_count dictionary, as a key. The corresponding value is a geometric mean of the number of word in each dictionary from the arguments. The specific words except preposition,

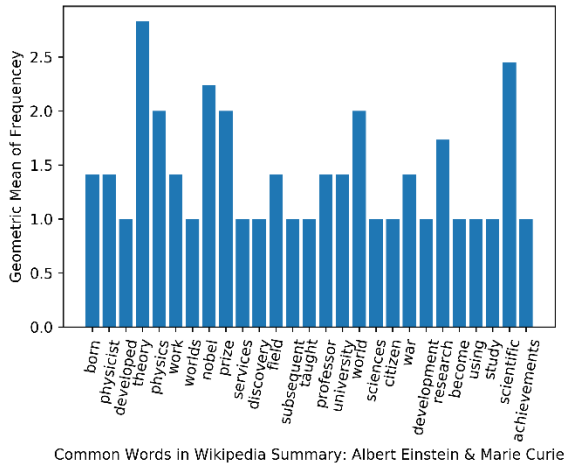
pronoun, number, article, common verb, alphabet and conjunction can be stored in the common_count dictionary.

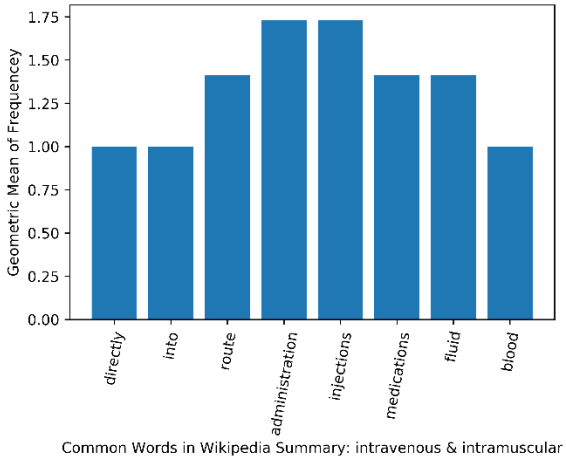
The last part of code is about taking word from user, and constructing a histogram based on the result. Dictionary of common word gets changed into histogram with designated bar, sticks, and label of each axis. Finally the histogram file is saved as .png extension.

Results

Several words are compared by using the program: 1) Olin College / Babson College, 2) Albert Einstein / Marie Curie, and 3) Intravenous / Intramuscular. Followings are each histogram obtained from the program.







From the result, it was able to find the common concepts between the given words. In the result of Olin & Babson, we can see that it is private college with undergraduate students in Massachusetts. Also in the comparison of Albert Einstein and Marie Curie, we can easily conclude that they are physicist, got Nobel prize, were professor in the university, did research, and made development and achievements. Also from the word “war” and “word”, it is expected that both of them are included in the world war. In the last case, the word intravenous and intramuscular are considered as difficult terminology, but the results help us to understand the words. It is related with the “directly”, “injections”, “fluid”, “blood”, and “route”. Now we can speculate what both words mean in common.

Reflection

I think I can improve the work into for multiple words, and I am expecting that I can make a full sentence which can summarize the long description based on this program.