Narratives for Final Project

**Investigation period:**

Before I actually start to work on this project, I set up several goals to help me select a suitable project. Firstly, I believe it is important to choose a project which is relevant to what I have learnt in this semester. It would be a chance for me to test myself whether I really understand the knowledge and whether I can implement them into a real problem, since most homework solutions could be found in the class demo. Hence, I initially planned to complete a little game, as it is more challenging. For example, I may need to install some new packages such as ‘pygame’ etc. Also, design game interface is also something I have not encountered yet.

However, in the later phase of my former project, I found it is hard to install the package on my computer. Also, the code for realizing the little game is much more complex than what I expected. Therefore, I decided to change my plan. After careful consideration, I planned to do a project related to text processing, which contains the knowledge of text cleaning, text extraction, dictionary, read files, Regex method, index, if-else and for loop.

Secondly, from my perspective, I should not only focus on the length of codes of the project, but the thoughts, investigation and execution of it. Therefore, I started to seek a text file which meets up with all of my requirements. At this moment, I came up with an idea that I can process a text file about vocabulary which contains the word itself, the part of speech, meaning as well as the example sentence. I planned to convert the vocabulary file into a structured dictionary, and then write it out to a csv file with clean format.

Thirdly, I want my project could provide some convenience for practical use. For example, what I have done could be used for similar situation and offer more convenience. Therefore, I believe convert vocabulary lists to a clearly formatted csv file would be very useful not only for me, but also those who may have to prepare for GRE/GMAT exams. So, for this time, I randomly chose a vocabulary list for SAT exam. The general thought is to let the word itself be the key in the dictionary, and the values are the part of speech, meaning and the example sentence. Also, I don’t want several separated lists of words. I plan to combined them into a full list, which means I need to update the serial numbers of them.

**Execution:**

In the phase of working on the code of the project, I encountered several challenges. Firstly, it is important to have a clear idea and plan about the whole project. However, at the beginning, it was hard for me to think clearly about what to do from the start to the end. How I solved it was to structure roughly at first, and added more detailed plans into it later. For example, my rough plan was to read the file into Pycharm and format all the lines to make them look clean and tidy. It does not have to be detailed, because when I got to work on the plan, I would realize the specific work to be done. For instances, I need to add new lines after full stops to format them, which was based on careful observation of the text file. Also, another detailed plan was to delete all the titles with key words of ‘SAT’ or ‘Vocabulary’. Then, I got a list of clean lines.

The main part of the project is to extract four parts out and assign the key and values. In this step, I met several difficulties.

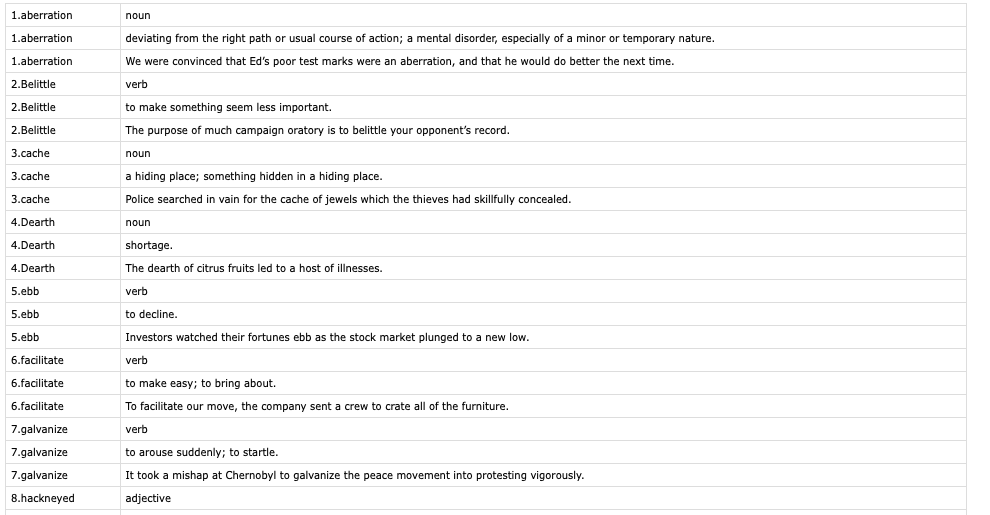
Firstly, I struggled about how to extract only the digits + a single dot + some words. Then, I realized I could use Regex, which was learnt in the later of this semester. I googled the string expression of ‘a number and a dot and a word’ and successfully got a list of numbers with the words followed.

Second challenge was to sequence the words. Initially, I planned to extract the numbers and modify them, then match them back with the corresponding word. However, it was hard to do so. So, I changed my plan to a new one: ignore the old serial numbers and use a counter to add a new series of number before words.

Another challenge was to slice different parts out. I managed to extract the part of speech by index the ‘-‘ character and the meaning of word is the part follows. However, when it comes to index the example sentences, some problems appeared. For example, when using the condition “elif ‘-‘ not in lines”, it may also return the result of the serial number as well as the words. However, if I add a condition of ‘if number not in lines’, this will cause another problem: in the example sentences, there are also numbers. Therefore, after carefully observing the data, I decided to use Regex again to match exactly a digit and a single dot. Also, there were some strange empty elements, which have not cleaned completely. Hence, I add them into the condition to delete all the empty elements.

Then, using the knowledge of dictionary to assign the key and values is not very challenging. Use ‘zip’ to zip all the values and generate the dictionary. Then, write out the dictionary to a csv file by using two for loop. The final csv file has the following format:

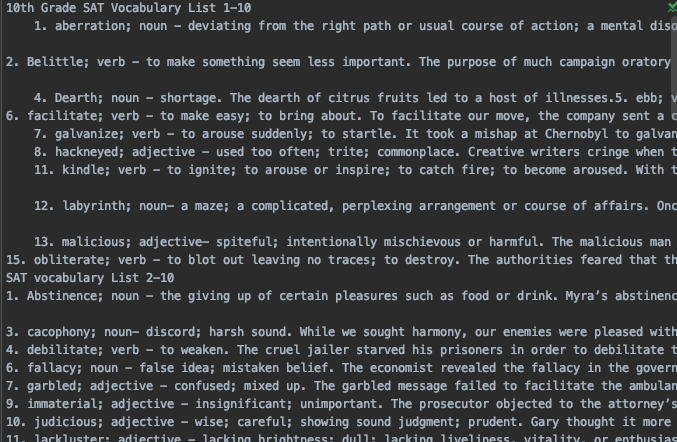
Figure 1 my version of the vocabulary file



**Reflection:**

1. Regex is really useful for matching and extracting elements.
2. Dictionary is good for structure, which allows me to re-organize the text file to a clearer format. (original: figure 2; my version: figure 1)

Figure 2 original version of the vocabulary file



1. It is very important to observe the data file before you actually start. This means you should observe the similarity of the data you want to process, i.e. what is the same point of all the example sentences? How are all the part of speech similar to each other? How to use several lines of codes to extract them out?

In summary, I believe this project gave me a valuable chance to reflect on myself. It combines the knowledge throughout the semester, and it is very practical. After completing this project, I believe I made some more progress on the processing of text files in Python. Furthermore, during the phase of convert the dictionary into a clear format of table, I learnt some methods by using pandas on Stack Overflow. I realized that the core of coding is not the length of codes, but the time complexity. When I did my mid-term project, I wrote like over 100 lines of codes and finally found it was too complex and a bit hard coding. Now, I no longer consider a longer chunk of code to be better, instead, complete the same task with shorter length of code is better. IS 452 is such a useful course to help me learn Python as a new bee. Thank you.