Trends in Word Usage--Adithya Shastry

Summary:

Project 8 was done to implement the priority queue data structure using an ArrayList. The priority queue was used to analyze reddit data and output the top 10 words used over a number of years of Reddit comments data. The priority queue was implemented in the PQHeap class which used an ArrayList to actually store the data. This PQheap was then used to make an ordered list of all of the common words in the reddit data based on the number of times the word was used. Once the data had been correctly sorted by the PQheap class. Finally, another method was added to check to see the frequency of words that the user entered. This was done in the word counter class, which is what reads in the comment data. This data was used to see if there were any trends in the words the user entered.The results were put into excel for analysis.

PQ Heap:

The PQ heap class was the class that was used to create the list of top ten words from the reddit file. This was done using an ArrayList and by changing the indexes of the items to make it so that index 1, I left out index 0 since it wouldn't make sense in terms of the calculations between the child and parent nodes, held the key-value pair with the highest count of words. Since it is very difficult to populate the PQ Heap directly from the reddit comment data, I first populated a hash table. This hash table then outputted a reddit comment output file with the word and the number of times it was used. This was then used to populate the PQHeap since I would not have to create a get method. Displaying the data was as simple as removing the first item from the heap 10 times and displaying it in the console with a print statement. The results can be seen in the results section below.

Frequency Counter:

The Frequency counter was a method added to the Word Counter Class. I decided to this because of the extra memory my program would need to allocate as I would need to create another class object to hold the Frequency Counter class. This was very easy to do since I could use a method I had already created to do this. The idea was to use the getKeyValue() method from the hash table class in order to find the word that the user intended to find. This was then outputted in a nice; easy to read format.

Extensions:

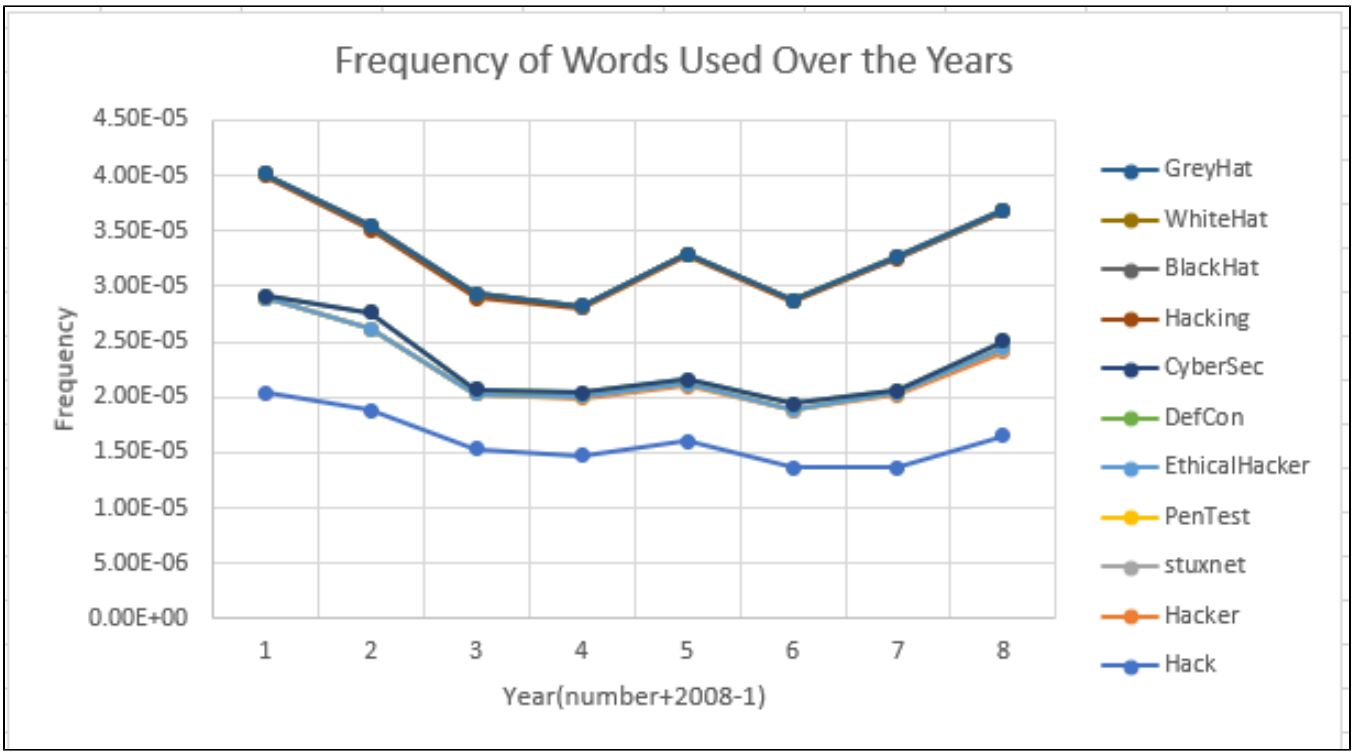
Interface to check for the frequency of words:

I created a nice interface that users can use in order to actually enter the words. This used the Scanner class from the Java API and allowed users to enter as many words as they would like and would display the data after every run. Thus I believe, makes it much easier for the user to actually input the words they need to and read the data that is outputted by the program.This can be seen in the results section below.

Using Special words:

As an aspiring entrepreneur, I decided to use words that pertain to the business I am currently working on. This business is based on Cybersecurity so in order to see where we are in terms of the discussion of cybersecurity topics, I decided to use many words that apply strictly to cybersecurity. I plan on using this to see if cybersecurity is actually a discussed topic on reddit. The words I used are below:

- **Hack**
- **Hacker**
- **stuxnet**
- **PenTest**
- **EthicalHacker**
- **DefCon**
- **CyberSec**
- **Hacking**
- **BlackHat**
- **WhiteHat**
- **GreyHat**

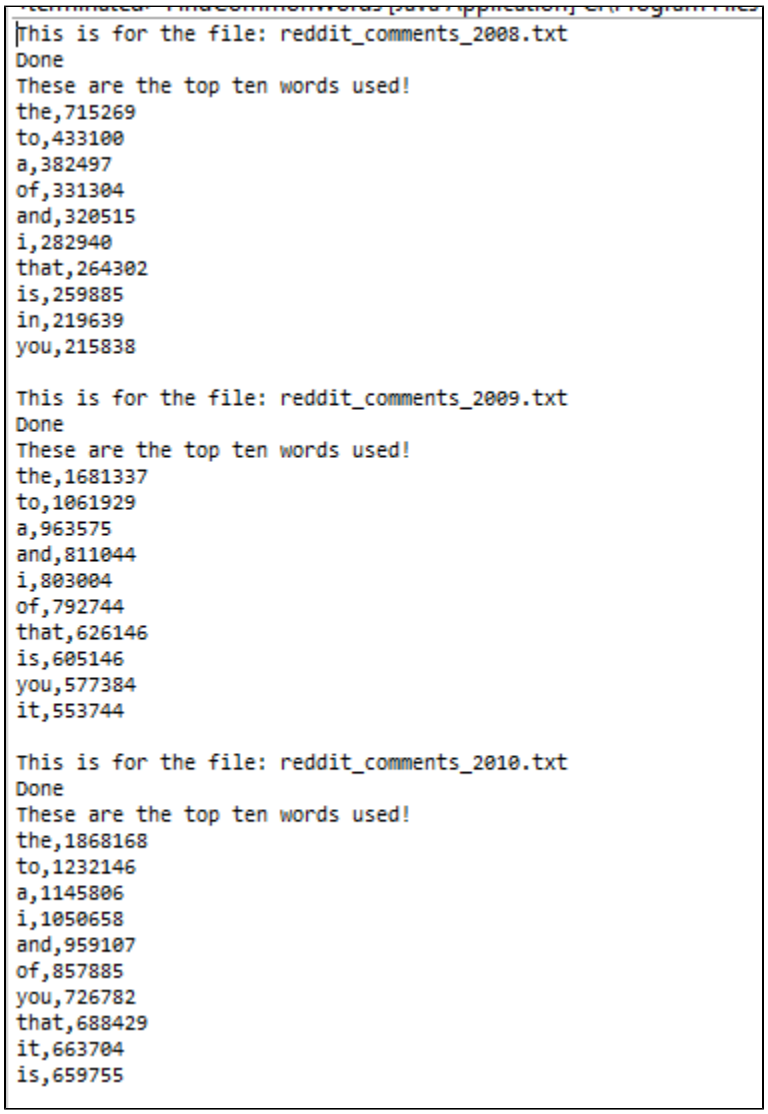


These words are very related to cybersecurity and thus are not used in any other form of language, at least to my knowledge but I could be wrong(English is my second language after all). This shows me that the topic of cybersecurity has gradually grown to a peak in 2015. This tells me that I should definitely continue to develop my cybersecurity product because there seems to be a great amount of discussion about the topics under cybersecurity. Most of these trends can be explained using the various events that have happened in the world of cybersecurity that has leaked to the popular media, but I still cannot explain why these topics were so important in 2008.

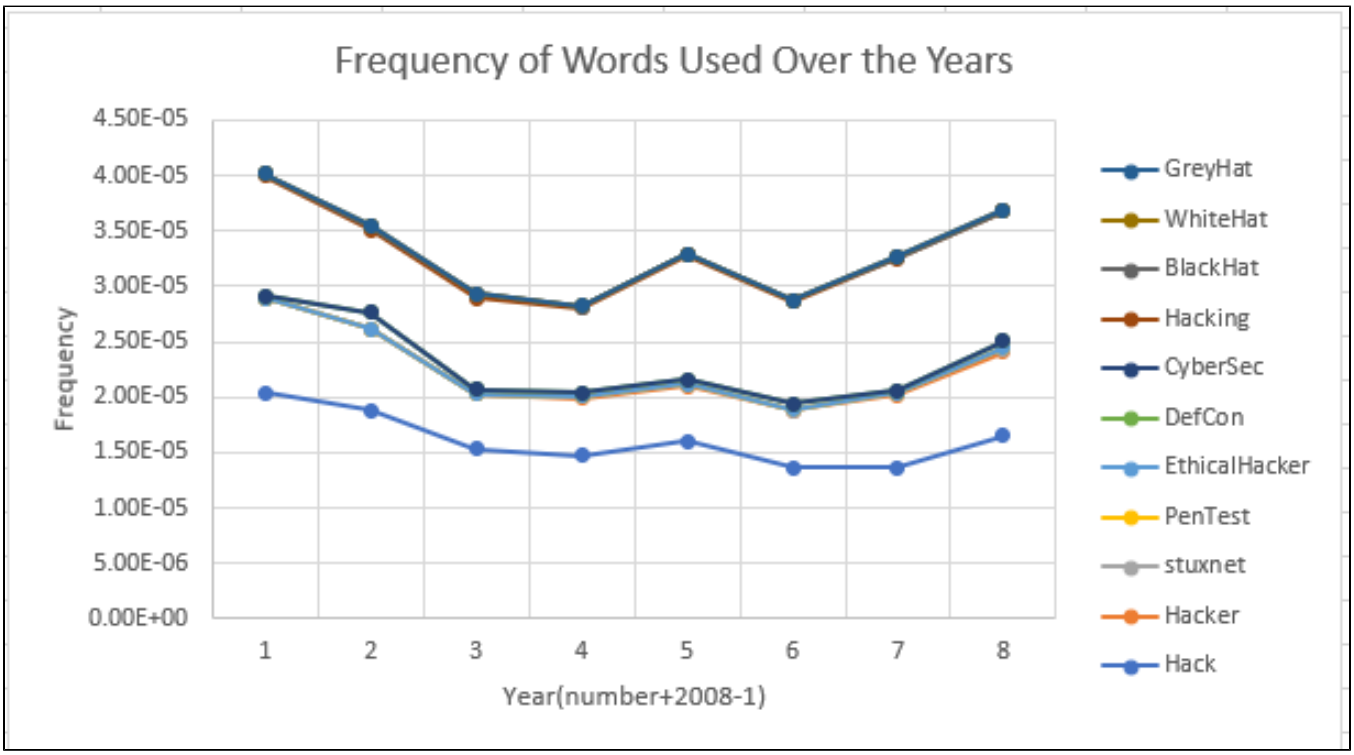
Results:

These are the Top ten words for the first three years:

This makes sense since these are words that have to be used in order for sentences to be gramatically correct.



The frequency of Chosen Words over the years:



Conclusion:

Overall, this was a very interesting project that gave me a lot of insight into a topic that I am very interested in!