Adrian Gellert

Professor Maximillian Bender

CS231 Lecture A, Lab D

18 December 2022

Adrian's Project 8: Trends Across the Sports Industry

ABSTRACT

In this project we analyzed Reddit word count files that we previously made in project 6. I also implemented methods that I created in project 7 to place String-Integer pairs from these files into a hashmap. One goal of the two analyses as part of project 8 was to find the ten most common words in each word count file to see whether common words and their frequency of use change over the years. The second goal of the two analyses was to determine whether a set of words under a common theme have trends in usage across years of Reddit word count files. To fulfill the first goal, we created a separate priority queue heap class to make the process of finding most common words faster than simply implementing binary search tree and hashmap methods from project 6 and 7. In contrast to these methods, the heap we implemented satisfied the rules that each parent node has a value larger than that of its children such that removing the root would remove the word with highest usage frequency. For the second goal, I simply implemented methods from project 7, and searched up famous, internationally recognized 21st century athletes.

RESULTS

Table 1: Top 10 Most Frequent Words per Year

Year	10 Most Frequent Words
2008	the, to, a, i, of, and, that, it, is, you
2009	i, to, a, and, it, of, you, that, is
2010	the, i, to, a, and, it, of, you, that, is
2011	the, i, to, a, and, it, of, you, that, is
2012	the, i, to, a, and, it, of, you, that, is
2013	the, i, to, a, and, it, of, you, that, is
2014	the, i, to, a, and, it, of, you, that, is
2015	the, i, to, a, and, it, of, you, that, is

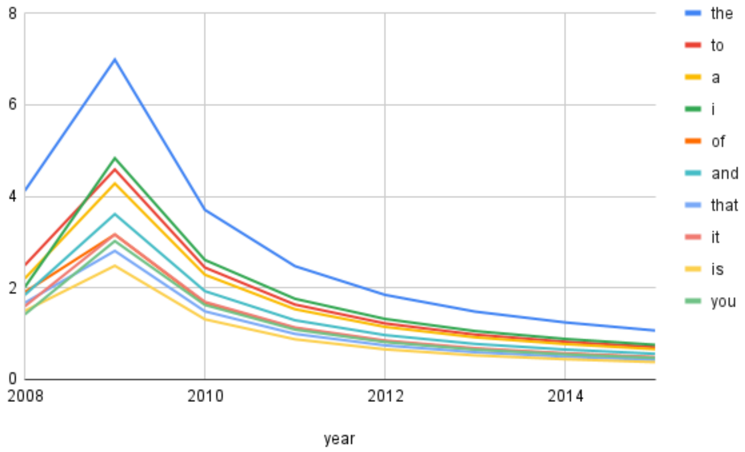

Figure 1: Change in frequency of 10 most common words (2008-2015)

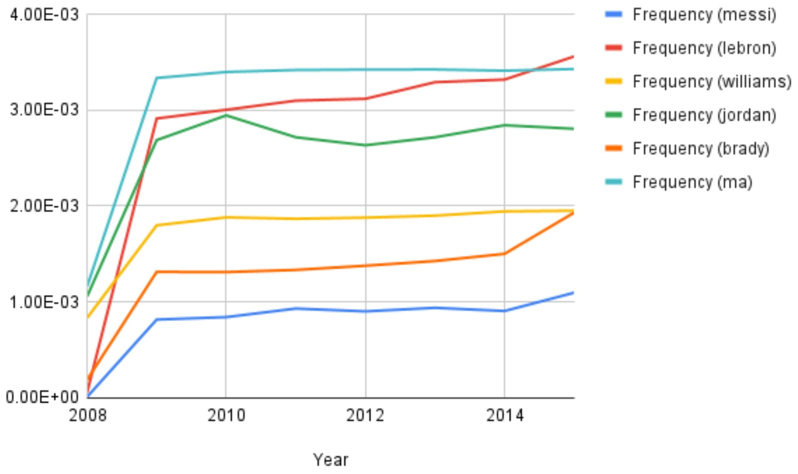

Figure 2: Change in word count frequencies for internationally recognized athletes

RESULTS DISCUSSION

As we can see from table 1, the most common words did not change between 2008-2015. However, from figure 1, we can see that the usage has a hump-shaped graph with a positive skew. I find it interesting that many of the words have a trajectory towards no usage. Possibly, this means that online language is going towards not using articles or pronouns.

From figure 2, we can see that the frequency of usage of words related to well known athletes is logistic, with a spike in usage during May 2009. Upon researching online, these spikes in usage make sense. For example, prior to the 2009 Italian Open, Serena Williams was “unseated from the top spot by Russia's Dinara Safina,” (The Guardian). As another example, in the 2008-2009 basketball season, the Cleveland Cavaliers had a record season in the NBA with 66 wins and 16 losses. In addition, LeBron James won his first MVP award. These events could have been the reason for the increased usage of their names.

REFLECTION

We used a priority queue heap in this project to find the most common words in reddit comment files. A priority queue heap is more ideal as a data structure choice than both a hashmap and a binary search tree because of time complexity differences. Although it takes the same time to remove an object and return it in both a binary search tree and a priority queue heap, O(log n), PQ heaps store highest value items at the root of the “tree” rather than down the length of the right side of the tree (as in binary search trees). Therefore, finding the highest priority value in a PQ heap is much faster O(1) time rather than O(log n) time. In comparison to hashmaps, PQHeaps are also much faster in this process of finding the highest priority word and its associated value. Hashmaps only make it faster to place and remove key-value pairs. However, there is no way of knowing which key-value pair has the highest priority in a hashmap without looping through all of the hashes in the map. Thus, finding a highest priority key-value pair becomes O(n) time even though removing it or peeking at it is O(1) time. To conclude, PQHeaps shorten the time to look for and look at a highest priority key-value pair, although removing that pair will have the same time complexity as a binary search tree (O(log n)), which is longer than the same method in a hashmap (O(1)).

COLLABORATION

I conversed with Jaime Yockey to understand differences between when the code for project 7 was similar to that of project 8 and when we should actually use methods from project 7.