CS231: Project 8

Word Trends

Parth Parth | CS231L-A | Dr. Alan Harper

**Abstract**

In this project, a Heap has been created to store words and their count in given files. A Heap is a tree-based data structure which is a complete (in this project, binary) tree that satisfies the “heap property” which is that any node in the heap is either greater than (if it is a max heap) or less than (if it is a min heap) than its parent node. The node at the top with no parents is called the root.

In the project, the heap has been used to analyze a. the ten most common words on Reddit between 2008-2015 and b. look at trends in data of words over the same time period.

**Analysis**

By running *CommonWordsFinder.java*, the following results were obtained:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
| 1 | the | the | the | the | the | the | the | the |
| 2 | to | to | to | to | to | to | to | to |
| 3 | a | a | a | a | a | a | a | a |
| 4 | of | and | i | i | i | i | i | i |
| 5 | and | i | and | and | and | and | and | and |
| 6 | i | of | of | of | of | of | of | of |
| 7 | that | that | you | you | you | you | you | you |
| 8 | is | is | that | that | it | it | it | it |
| 9 | in | you | it | it | that | that | is | is |
| 10 | you | it | is | is | is | is | that | that |

From analyzing the table, it was found that the most common words for all eight years were

1. the
2. to
3. a
4. i
5. and
6. of
7. you
8. it
9. is
10. that
11. except for 2008 when “it” was replaced by “in.”
12. “the” and “to” remained the top two most popular words with “i,” “and,” “of,” and “you” being fourth through seventh place starting 2010.

**Extension 0 / More Analysis**

For this extension, I have analyzed trends of custom words and plotted my results:

1. harvard, yale, mit, stanford, nyu, colby

Chart, line chart

Description automatically generated

My hypothesis would be that the initial dip seen between 2008 and 2012 is the result of the 2008 financial crisis when it became unaffordable for a lot of people to go to college. However, as the job market picked back up in 2012, people started going to college again, which explains the frequency stabilizing (and even reversing for MIT and Harvard). The 2013 uptick in Colby’s searches might be related to the scandal that involved the former Barclays top executive in late 2012.

1. expedia, amazon, google, microsoft, clearwire, dendreon

Chart, line chart

Description automatically generated

These are the most popular tech stocks in 2012. As can be seen, Google, Amazon, and Microsoft are much more popular than the other three stocks. While the decrease in Google could just be explained by the fact that the number of other words on Reddit increased, the increase in Microsoft and Amazon coincide with the release of Windows 8 (Oct 2012 which shows in 2013) and the stock market heavily betting on Amazon in 2013 (after Amazon stock started increasing rapidly).

**Extension 1**

In this extension, I have implemented a new class *CommonWordsFinderArrayList.java* that uses an ArrayList to store words. It is then sorted and the top ten words retrieved. A run-time analysis of this process using a heap and an ArrayList has been done below.

|  |  |  |
| --- | --- | --- |
|  | Heap | ArrayList |
| 2008 | 0.0031345 | 0.0646494 |
| 2009 | 0.0038787 | 0.0782093 |
| 2010 | 0.0027109 | 0.0771274 |
| 2011 | 0.0028156 | 0.0869174 |
| 2012 | 0.0028086 | 0.0579286 |
| 2013 | 0.0036631 | 0.0485204 |
| 2014 | 0.0052887 | 0.0576521 |
| 2015 | 0.0040186 | 0.0686786 |
| Average | 0.0035398 | 0.0674604 |

As is clear from the data, the ArrayList is almost 20 times slower when compared to the Heap.

**References / Acknowledgements**

I took no help with the code from any source. The top stocks of 2012 were retrieved from <https://www.geekwire.com/2013/nasdaq-top-performers-expedia-takes-flight/> .