Text Analyzer – Project Outline

**Overview**

A text analyzer able to parse a given HTML file and output basic statistics about the text used in the file. Specifically, the frequency of all words used in the file sorted from greatest to least. For example, if the most used word is “the” and it is used 50 times, the output should be “the: 50” at the top of the list. The output should display to the Java console. Parsing should begin at the start of title header while ignoring all HTML tags when calculating the output.

**Scope**

No scope has been formally specified. Ideally the analyzer should be able to parse any HTML document to perform this task. However, the project will be deemed successful with the completion of this single document ([The Project Gutenberg E-text of The Raven, by Edgar Allan Poe](https://www.gutenberg.org/files/1065/1065-h/1065-h.htm)).

**Input**

* HTML file (or URL?) of indeterminant length

**Output**

* Sorted list of most frequently used words, sorted greatest to least, alongside the number of times each word appears in the text
* This list will ignore HTML tags and only begin at the title of the work and stop at the end of the poem

**Problems**

* How to get the parser to start only on the title line?
  + In the provided input file the title is given in an H1 header tag; no other text uses this tag until this point; start the parser after it reads “<H1 … >”?
  + This could impact the below HTML tag ignoring solution
* How to ignore HTML tags?
  + Do not parse text after a “<” character is read and continue to parse after a “>” character is read.
  + This could cause issues with document that include these characters in their text not as HTML tags.
* How to ignore HTML special character calls that are not bracketed such as “&mdash;”?
  + “&mdash;” seems to be the only code used in the body of this text; could simply create an exception for it but that limits application outside of this specific HTML file
  + Could build a database of exception character codes to use; this would affect performance and is outside the scope of this particular project; perhaps a smaller ArrayList instead?
* How to ignore the license text at the bottom?
  + The full poem is contained mainly with <P> tags as well as <H#> tags; set the parser to ignore any tags that are not of this type?
  + If the parser reads “<H” or “<P”, start logging words after the next “>” character is parsed.

**Project Notes**

* jsoup can be used as an HTML parser