1. Install Python (if you don’t have it already), and install NLTK.
   * Using Anaconda Cloud for Jupyter Notebooks I can import the nltk library as it is a part of the package
2. Follow the instructions in chapter 1 of Bird-Klein for implementing a “lexical diversity” scoring routine.
   * See submitted J.Coate.NLP.Unit1.Homework.ipynb
3. Go to <http://www.gutenberg.org/wiki/Children%27s_Instructional_Books_(Bookshelf)>, and obtain three texts (of different grade levels) from the “Graded Readers” section. Report the lexical diversity score of each. Explain whether the result was surprising.

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Book** | **Grade** | **Lexical Diversity Score** |
| Lewis Carroll | Through the Looking Glass | 0-3 | 10.633970697789918 |
| Mark Twain | The Adventures of Tom Sawyer | 6-8 | 9.854816093193831 |
| Rudyard Kipling | The Jungle Book | 1-4 | 10.65876389797253 |

* + These results are surprising because they are very close together, despite being different grade levels. I may have picked ranges of grades too close together to properly leverage this function.
  + See submitted J.Coate.NLP.Unit1.Homework.ipynb for code on how the above was obtained

1. Also compare the vocabulary size of the same three texts. Explain whether the result was surprising.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Author** | **Book** | **Grade** | **Lexical Diversity Score** | **Vocabulary Size** |
| Lewis Carroll | Through the Looking Glass | 0-3 | 10.633970697789918 | 4027 |
| Mark Twain | The Adventures of Tom Sawyer | 6-8 | 9.854816093193831 | 9271 |
| Rudyard Kipling | The Jungle Book | 1-4 | 10.65876389797253 | 6116 |

* These results are a little more along the lines of what I would expect to see. For instance as the children get older they are able to handle books that are longer and contain more unique vocabulary as whole.

1. Write a paragraph arguing whether vocabulary size and lexical diversity in combination could be a better measure of text difficulty (or reading level) than either measure is by itself.
   * With the above information it become apparent that grade ranges for the difficulty of a book vary wildly. We can see that the lexical diversity between these books is quite close. Indicating that there is not a lot of variation between them and therefore, theoretically, could all be handled by the same age range. However, the length of these books vary quite a bit. Therefore, the only thing that would be really challenging for the student/child would be length of time or attention span needed to read the assignments. This being stated, I can confidently state the evidence suggests a more accurate measurement of book level and reading level for books would be using a combination of the lexical diversity score alongside the vocabulary size to determine a proper text difficulty measurement for students moving forward.