## Julia Layne - Homework 1

1. Install Python (if you don’t have it already), and install NLTK.
2. Follow the instructions in chapter 1 of Bird-Klein for implementing a “lexical diversity” scoring routine.
   1. <http://www.nltk.org/book/ch01.html>
   2. Section 1.4
3. Go to <http://www.gutenberg.org/wiki/Children%27s_Instructional_Books_(Bookshelf)>, <http://www.gutenberg.org/ebooks/bookshelf/215> 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.

def lexical\_diversity\_words(text):

text = text.split()

print("Set size: ")

print(len(set(text)))

print("Length of entire text: ")

print(len(text))

return len(set(text)) / len(text)

* 1. I was surprised to see that the lexical diversity score seemed to go up and then back down. I had expected them to g

|  |  |  |
| --- | --- | --- |
| Level | Name | Lexical Diversity Score |
| 1 | McGuffey's First Eclectic Reader by William Holmes McGuffey | 0.1977 |
| 3 | McGuffey's Third Eclectic Reader by William Holmes McGuffey | 0.2309 |
| 5 | McGuffey's Fifth Eclectic Reader by William Holmes McGuffey | 0.2145 |

1. Also compare the vocabulary size of the same three texts. Explain whether the result was surprising.
   1. Looking further into the set size and total length, it appears that the set size of the text does go up, while the total number of words grows at a greater rate. This lengthening of the text in the Level 5 text is what is really pulling down the diversity score back down.

|  |  |  |
| --- | --- | --- |
| Level | Set Length | Words split in Total Text |
| 1 | 2010 | 10165 |
| 3 | 7078 | 30648 |
| 5 | 21827 | 101749 |

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.
   1. The vocabulary size appears to be a better ranking of the text levels. When looking at lexical diversity, this should be used in combination with the lexical diversity to get a fuller idea of the text. Lexical diversity is a function of the length of text. This can create the problem where two similarly complex in vocabulary books would have different lexical diversity. Particularly for reading level this is worse, because a longer book will be judged as less diverse and ‘easier’ than a shorter book with similar vocabulary. The kid having to read the book would beg to differ. Reading a complex vocabulary 5,000 word short story is likely going to be much easier than the 200,000 word book using the same set of words. Lexical diversity is best kept for comparing texts with similar lengths.