## 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.
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.

* The following is the lexical diversity scores for each of the books I checked.
  + The Bacon Second Reader
    - lexical diversity = .06811 %
  + Mcguffy’s Forth Eclectic Reader
    - lexical diversity = .02301 %
  + Mcguffy’s Fifth Eclectic Reader
    - lexical diversity = .01043
  + The Ontario High School Reader
    - lexical diversity = .02638
* The suprising thig here is that the book with the highest lexical diversity is the second grade reader (or the least advanced book). This may be because the least advanced book is simply shorter then the other books, and therefore has a less significant lexical diversity. In fact, in the books listed, each subsequent grade, with the exception of high school, has a lower lexical diversity then the grades below it.

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

* Vocabulary counts for all books
  + The Bacon Second Reader
    - Unique word count = 81
  + Mcguffy’s Forth Eclectic Reader
    - Unique word count = 87
  + Mcguffy’s Fifth Eclectic Reader
    - Unique word count = 84
  + The Ontario High School Reader
    - Unique word count = 127
* In the case of vocabulary counts, the result is more in line with expectations. This shows that higher reading level books have a larger vocabulary. The only exception here is the fourth grade to fifth grade transition; where fifth grade has slightly less overall unique words then fourth.

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.

* Based on the data collected in this assignment, Vocabulary size seems to be a better measure of reading level then lexical diversity, but even with this measure, there is still an example that diverges from the trend. Perhaps the best method of measuring text reading difficulty would not be to measure the overall word count or lexical diversity, but to measure the frequency of word appearance in relation to how often the words appear in all English language. It also seems probable that it is not easy to measure reading difficulty by measuring a single variable. It would have to be placed within a larger context. For example, if one book has a very large vocabulary of commonly used words; while another book contains more ecelectic language, but contains overall less vocabulary, the later book would likely be the more difficult one to read.

Appendix (python code)

```{python}

*from nltk.book import \**

*import re*

*start\_line = '\\*\\*\\* START OF THIS PROJECT GUTENBERG EBOOK.\*\\*\\*\\*'*

*end\_line = '\\*\\*\\* END OF THIS PROJECT GUTENBERG EBOOK.\*\\*\\*\\*'*

*## Functions*

*def lexical\_diversity(text):*

*if text is str:*

*text = split\_text(text)*

*return 100 \* len(set(text)) / len(text)*

*def vocab\_score(text):*

*if text is str:*

*text = split\_text(text)*

*return len(set(text))*

*def split\_text(text):*

*return re.split("[^a-z']+")*

*def remove\_meta(text):*

*return re.split(end\_line, re.split(start\_line, text)[1])[0]*

*### Output*

*grade2 = open('data/beacon\_second\_reader.txt').read()*

*grade4 = open('data/mcguffys\_fourth\_reader.txt').read()*

*grade5 = open('data/mcguffys\_fifth\_reader.txt').read()*

*gradehs = open('data/ontario\_highschool\_reader.txt').read()*

*print(lexical\_diversity(remove\_meta(grade2)))*

*print(vocab\_score(remove\_meta(grade2)))*

*print(lexical\_diversity(remove\_meta(grade4)))*

*print(vocab\_score(remove\_meta(grade4)))*

*print(lexical\_diversity(remove\_meta(grade5)))*

*print(vocab\_score(remove\_meta(grade5)))*

*print(lexical\_diversity(remove\_meta(gradehs)))*

*print(vocab\_score(remove\_meta(gradehs)))*

*```*