# Apple Assignment

Analysis of a large body of text  
  
Download a plain text (UTF-8) version of Tolstoy’s War and Peace (<http://www.gutenberg.org/ebooks/2600>)  
  
Using Python, parse the text file into a structured format, then analyze the dataset and visualize features defined below.  
  
1. Write a parser that separates the body of the book's text from the header (including table of contents and Project Gutenberg license paragraph, body of the books starts with "BOOK ONE: 1805") and footer (starting with "End of the Project Gutenberg EBook”)  
  
2. Write a tool to structure the extracted body of text into the following nested format:  
  
Book number and year  
-> Chapter index (within a given book)  
-> Paragraph index (within a given chapter)  
-> Sentence index (within a given paragraph)  
-> Sentence text  
-> Word index (within a given sentence)  
-> Word (w/o special characters) (only alphanumeric values)  
  
Choose an appropriate container to store the nested dataset (e.g. dict or pandas dataframe).  
Choose an appropriate and efficient method to serialize the nested dataset into a file.  
  
3. Given this dataset, provide the following analysis:  
  
a. Ratio of direct (or quoted text) speech to indirect speech (or unquoted text, excluding titles and headers)  
b. Illustrate, with a simple and comprehensible visualization, the lengths of all books and chapters in one graph. (Optional: find/point out interesting features or outliers in the text structure)  
c. Perform a word frequency analysis, returning a table of word occurrences and provide a visualization  
  
The result you should return back to us should be:  
1. Documented python code (including all parser, serialization and analysis functionality) - can be a combination of .py files and/or jupyter notebooks  
2. A representative PDF slide deck (think presenting it to a group of Apple engineers) with plots and summaries of points a. b. and c. Feel free to include your work/thought process in the deck.