CSC 143 Programming Project #2

Text Analysis using Java Collections Framework

There is some debate on influence of Jane Austen on Charlotte Bronte work as a writer (and in general on all three Bronte sisters'). If you are interested in finding more, feel free to google for their works and the debate.

For this exercise, using the publicly available books on Project Gutenberg (http://www.gutenberg.org), you are asked to find the **top 10 words and number of times they occur in books by Charlotte Bronte but not used by Jane Austen**.

We will use the following books:

Jane Austen	Charlotte Bronte
Pride and Prejudice (http://www.gutenberg.org/files/1342/1342-0.txt)	Jane Eyre: An Autobiography (http://www.gutenberg.org/cache/epub/1260/pg1260.txt)
Emma (http://www.gutenberg.org/files/158/158-0.txt)	Villette (http://www.gutenberg.org/cache/epub/9182/pg9182.txt)
Sense and Sensibility (http://www.gutenberg.org/cache/epub/161/pg161.txt)	Shirley (http://www.gutenberg.org/files/30486/30486-0.txt)
Persuasion (http://www.gutenberg.org/cache/epub/105/pg105.txt)	The Professor (http://www.gutenberg.org/files/1028/1028-0.txt)
Mansfield Park (http://www.gutenberg.org/files/141/141-0.txt)	

HashMap (or HashTree) Java Collection will come handy for finding and keeping the count of words. You are only allowed to use Java Collections as described in Chapter 11 of our class textbook.

Hint: refer to Word Count Map case study in Chapter 11.

To receive full credit, submit the following:

- Java source code file(s)

- Nicely formatted report containing a **table of top 10 words found in Charlotte Bronte's books and *not* in Jane Austen's and their counts**, and a **summary of your insights** from this exercise including how this type of analysis can be useful and any suggestions for improvements.

- **Don't** submit the text files for the books!