Comprehensive Practice

Section One

Goal

- In this section we will develop a program that will read two different text files and compare t heir vocabulary.
- In particular, we will determine the set of word s used in each file and compute the overlap bet ween them.
- Researchers in the humanities often perform s uch comparisons of vocabulary in selections of text to answer questions like, “Did Christopher Marlowe actually write Shakespeare’s plays?”

- We will develop the program in stages:
- 1. The first version will read the two files and rep ort the unique words in each.
  - We will use short testing files for this stage.
- 2. The second version will also compute the over lap between the two files (i.e., the set of words t hat appear in both files).
  - We will continue to use short testing files for this stage.
- 3. The third version will read from large text files and will perform some analysis of the results
  - including the number of words in each list, the n umber of words of overlap, and the percentage of overlap.

Possible Solutions(1)

Stage 1

```
list = new empty list.
while (more words to process) {
    word = next word from file.
    if (list does not contain word) {
        add word to list.
    }
}
```

Possible Solutions(2)

- Stage 2

```
overlap = new empty list.
for (each word in list1) {
    if (word is in list2) {
        add word to overlap.
    }
}
```