# lab Assignment 20.3

# Countwords

**Background:**

1. This lab assignment will count the occurrences of words in a text file. Here are some special cases that you must take into account:

|  |  |
| --- | --- |
| **Special Cases** | **Explanation** |
| hyphenated words(i.e., *sixty-three*) | Count as one word |
| hyphenated words with blank spaces on each side of hyphen (i.e., *joyous – sparkling*) | Count as two words |
| apostrophed words (i.e., *'tis,* or *can't*) | Count as one word |
| upper and lower case (i.e., *The* and *the*) | Both count as occurrences of the word '*the*'. Convert any capital letters to lower case before counting such words. |

2. You are encouraged to use a combination of all the programming tools you have learned so far, such as:

|  |  |
| --- | --- |
| **Data Structures** | **Algorithms** |
| Array classes  String class  ArrayList class | sorting  searches  text file processing |

**Assignment:**

1. Your instructor will provide you with a data file (such as *test.txt*, *Lincoln.txt*, or *dream.txt*) to analyze. Parse the file and print out the following statistical results:

– Total number of unique words used in the file.

– Total number of words in a file.

– The top 30 words which occur the most frequently, sorted in descending order by count.

For example:

1 103 the

2 97 of

3 59 to

4 43 and

5 36 a

6 32 be

7 32 we

8 26 will

9 24 that

10 21 is

... rest of top 30 words ...

Number of words used = 525

Total # of words = 1577