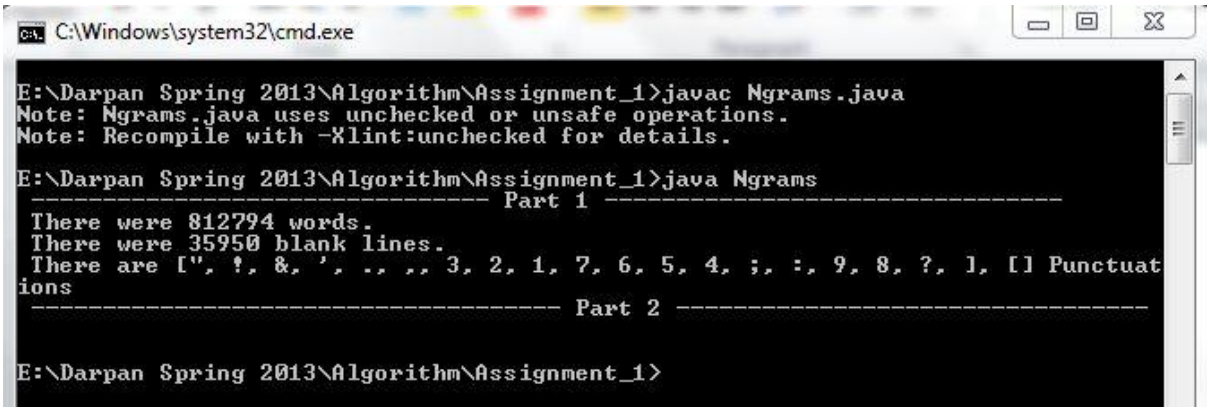
PROBLEM SET 1

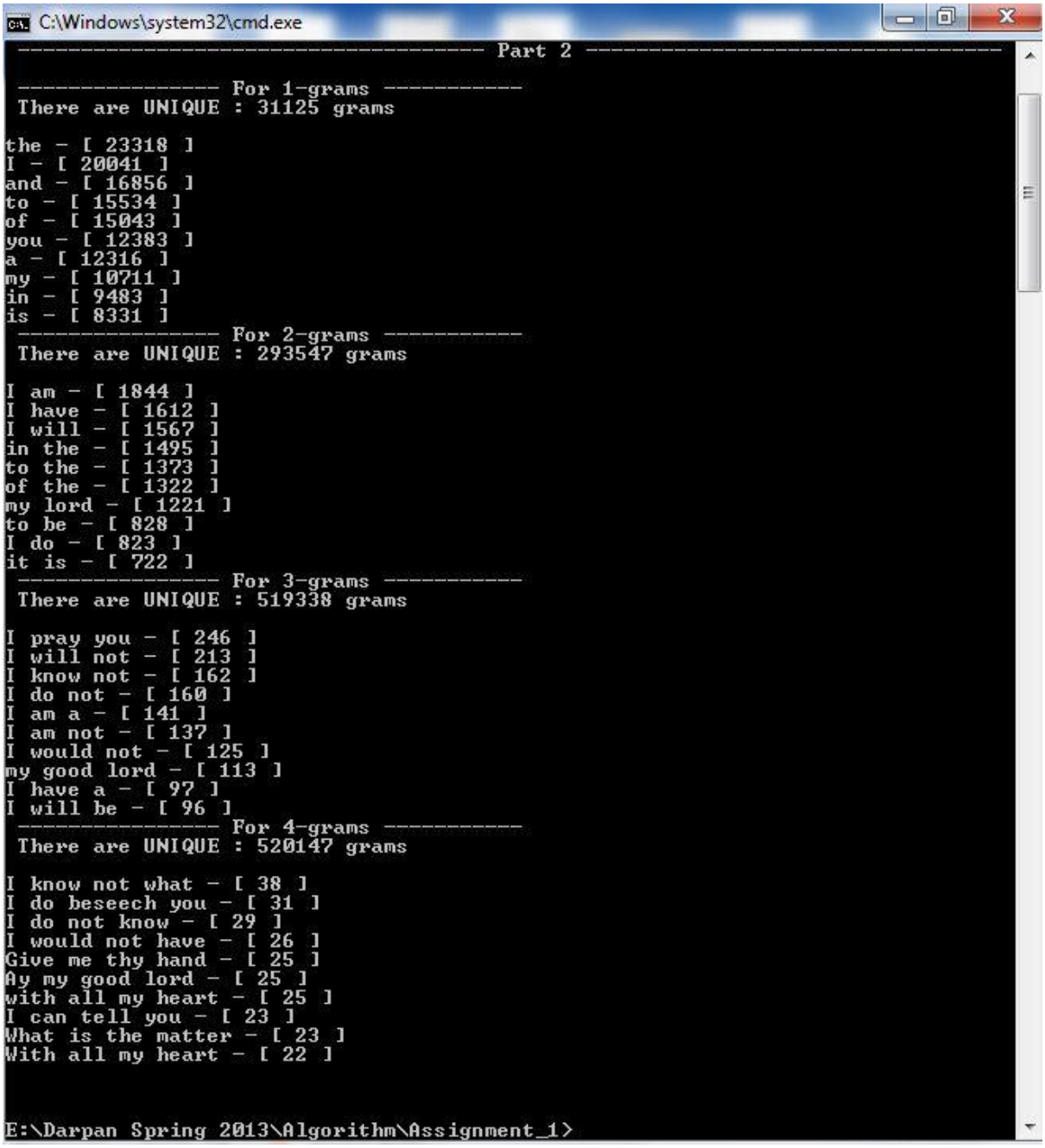
PART – 1:

- In the first part of the assignment, this program reads a line one by one using BufferedReader and then scans that each of the lines. Then it converts into the number of tokens of that line with the help of split().
- With the help of Regex, tokens are scattered into the words, special characters and count those words and characters and store them into the List.
- At the end of the file, our whole List will contain all the words. This List can be transferred to the SortedList so that we will get unique words. Meanwhile, this program also takes care for the special characters, and again store them into one another List and convert them into the SortedList, plus, it also counts the number of blank lines appeared at each and every line.
- Finally, it just prints all the SortedList and shows the special characters.
- For Example :



PART -2:

- In the second part of the assignment, first of all this program converts each of the line into the number of tokens and store them into the List. After that if there are special characters appear or the word is in the Upper Case, then it filters the special characters and convert them into the Lower Case with the help of String.lowerCase() method.
- This program uses the Map Data Structure to count the number of grams. Each and every line, it updates its value, this Map is like Map<String,Integer> , in which its key is tokens and its values are the number of repetitions.
- Once the whole file has been finished, we get the Map of whole tokens. For n-grams this program passes this map to the function “ngrams()” as a parameter and this ngrams returns the List of n-grams.
- With the help of Comparable Interface, it has used the comapreTo() method to sort them according to the values of Map. And it gives the expected output.
- For Example :



Normally, This program takes 2-3 seconds to finish the part 1. And for the part 2 it takes probably around 10-15 seconds to complete the process.