Q1

|  |  |
| --- | --- |
| **WSIZE** | **# duplicates** |
| 2 | 60,453 |
| 3 | 10,756 |
| 4 | 1,987 |
| 5 | 667 |
| 6 | 362 |
| 7 | 226 |
| 8 | 151 |
| 9 | 105 |

Q2

1. The results of these methods should be the same because in both methods, all items are iterated through and then added to the HashSet, which should filter out all the duplicates.
2. Conceptually the two benchmarks are different because the first benchmark iterates through the words and then adds WordGram objects with sequential possibilities of lists whereas the second benchmark iterates through the words, adds a WordGram current with the first possibility and then shifts current to include the next possible word and then adds that WordGram.

Q3

Algorithms most definitely can be biased. The data sets used to train machine learning often do not accurately represent certain populations of people and because of that, they may not accurately determine results and conclusions for those populations. However, because stratifying data sets to understand different populations may also have underlying discriminatory intent or bias, finding ways to train machine learning data is also difficult and will still result in bias.