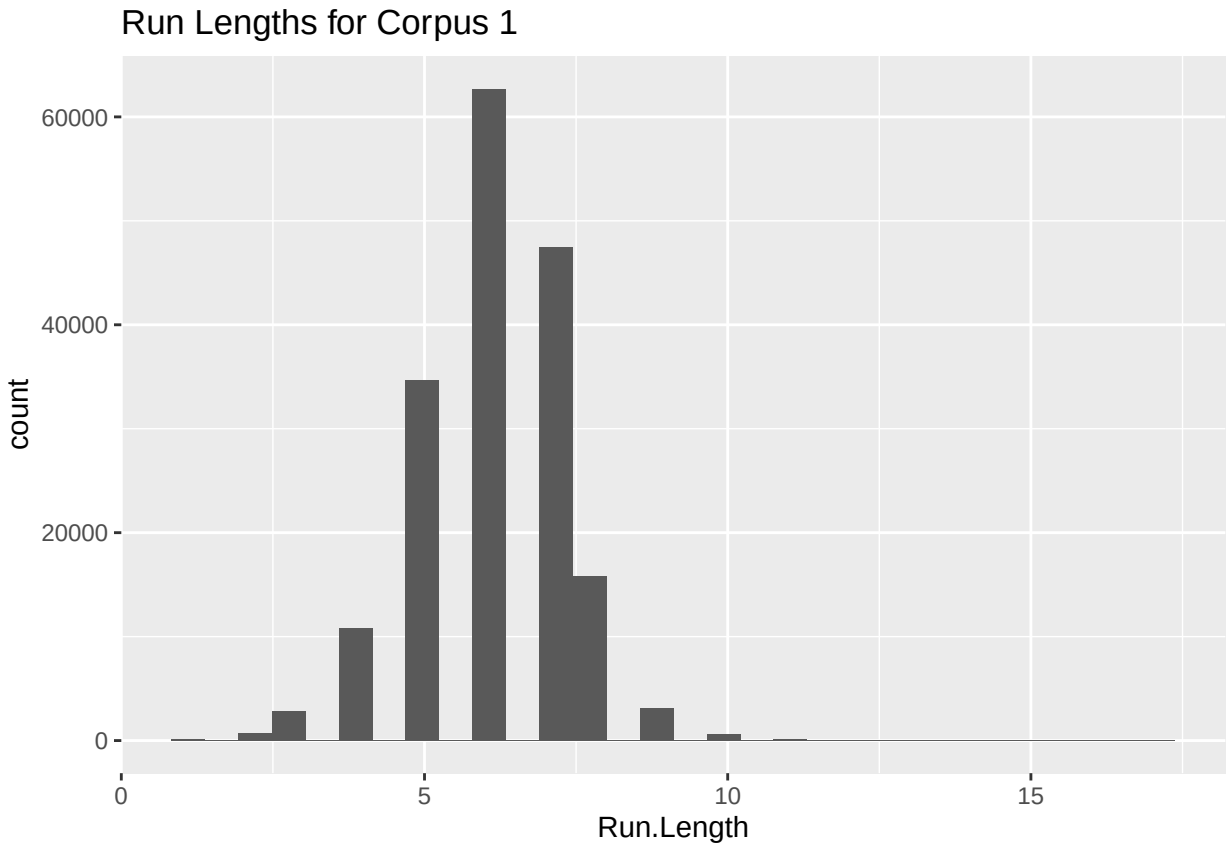
Run Distrubution

Caden Corontzos

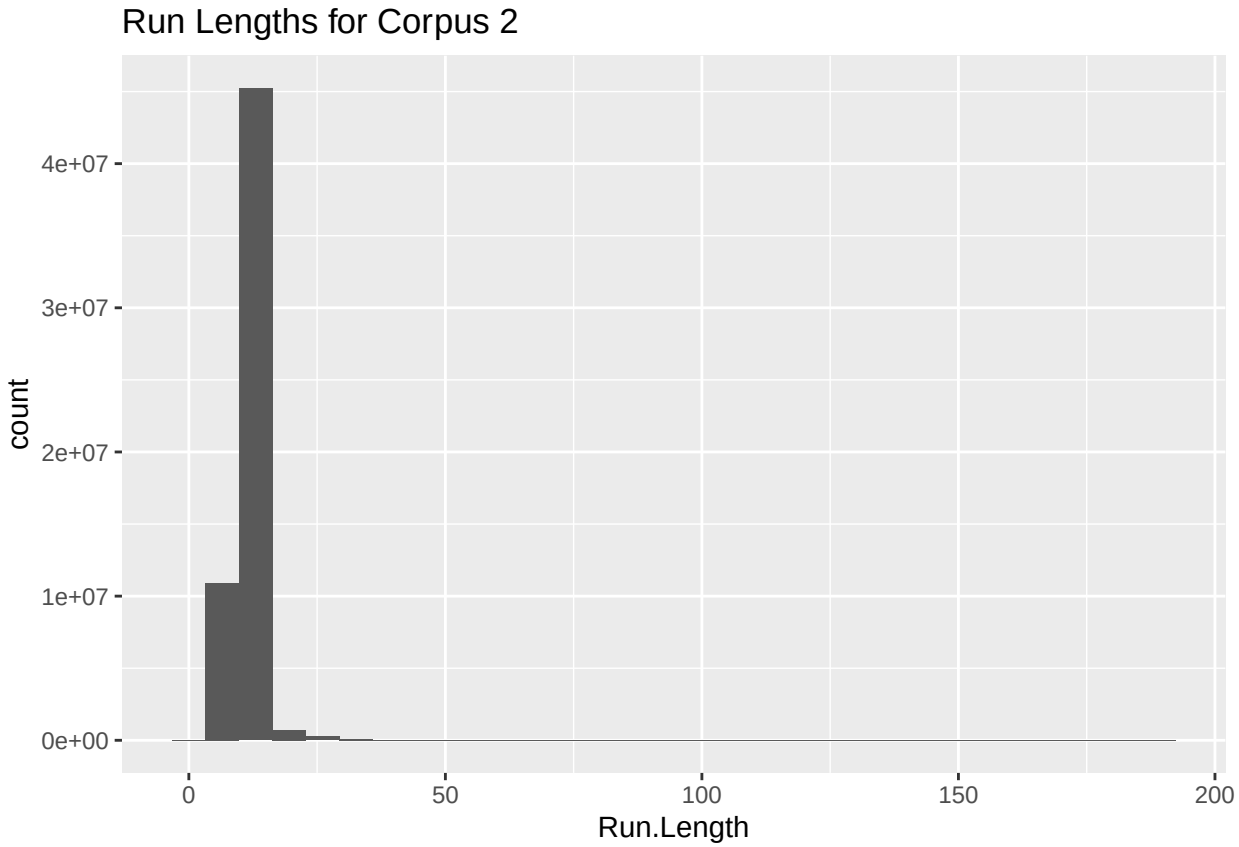
On each iteration of LZW, the algorithm keeps adding letters to a string until it finds the longest string that it has already seen before. If we know the distribution of the lengths of these strings, we may be able to better optimize our dictionary data structure.

I ran the basic LZW algorithm on the two corpora and counted the length of these runs.



A summary for Corpus1.

```
##   average_run_length maximum_run_length median_run_length sd_run_length
## 1           6.135398                 17                 6      1.237217
```



A summary for corpus 2

```
##   average_run_length maximum_run_length median_run_length sd_run_length
## 1           10.96825                190                11      2.494305
```

A summary of both corpora.

```
##   average_run_length maximum_run_length median_run_length sd_run_length
## 1           10.95318                190                11      2.505904
```