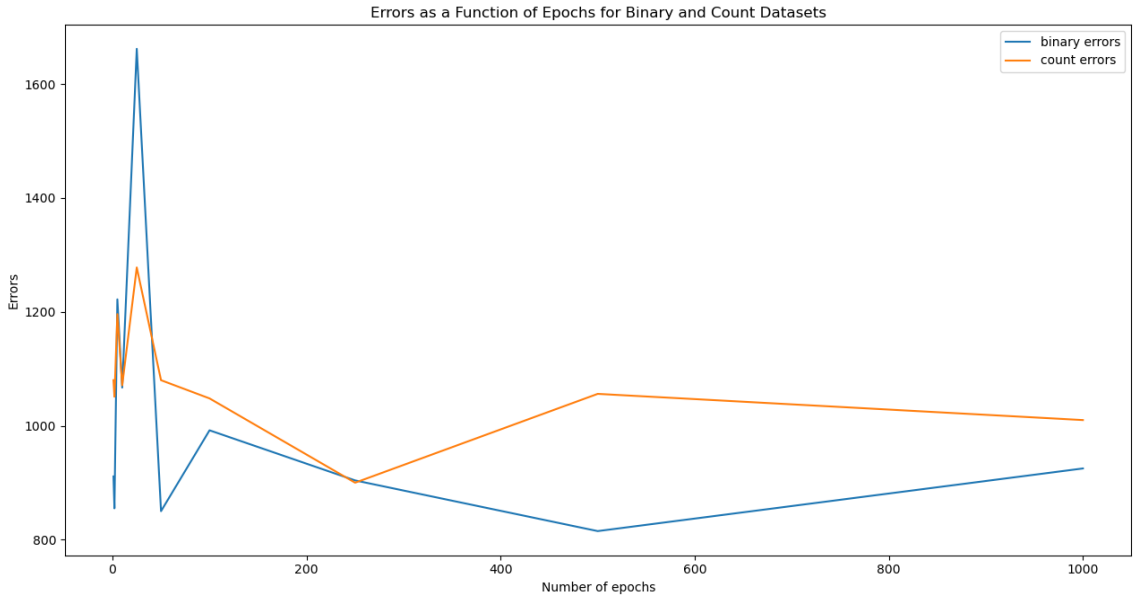
/* THIS CODE IS MY OWN WORK, IT WAS WRITTEN WITHOUT CONSULTING CODE WRITTEN BY OTHER STUDENTS OR LARGE LANGUAGE MODELS SUCH AS CHATGPT.
Tommy Skodje */

I collaborated with the following classmates for this homework:
None

2. c.

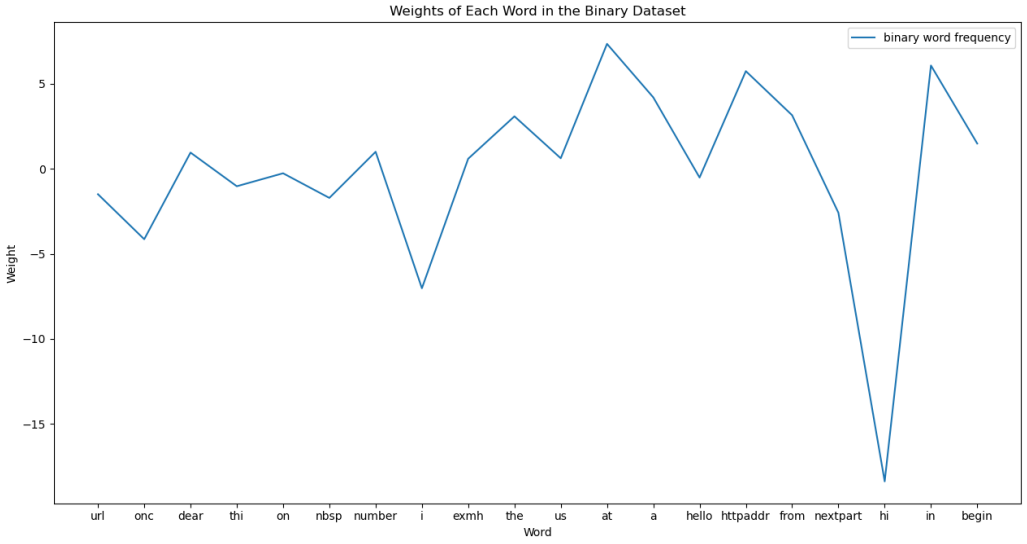


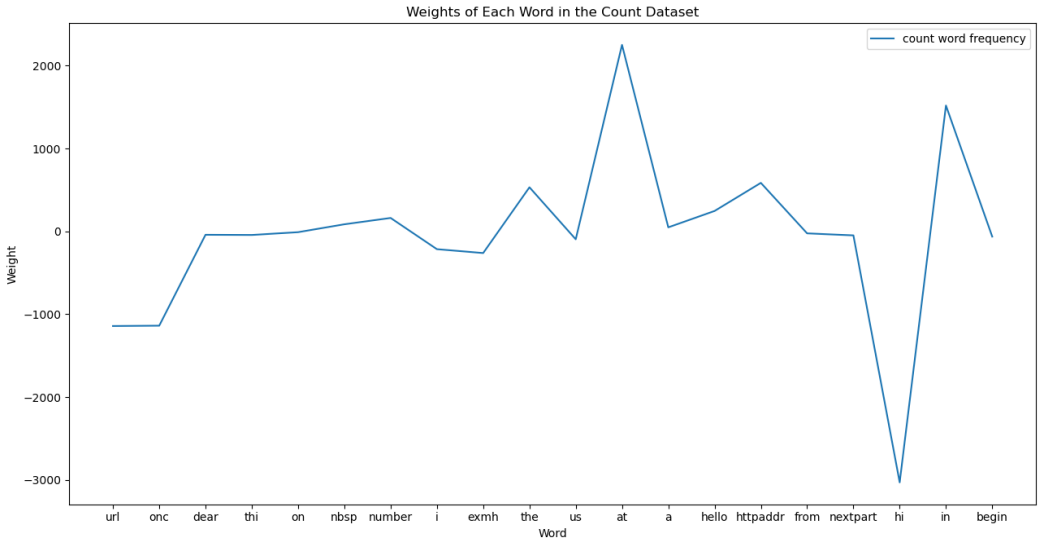
As shown in the graph above, the optimal number of epochs for the binary dataset was 500, and the optimal number of epochs for the count dataset was 250.

```
binary training mistakes: 881
binary test mistakes: 376
count training mistakes: 933
count test mistakes: 424
```

The above screenshot shows the results of training and predicting a perceptron with 500 epochs for the binary dataset and 250 epochs for the count dataset.
The number of mistakes the binary model made on the training dataset was 881, and the number of mistakes it made on the test dataset was 376.
The number of mistakes the count model made on the training dataset was 933, and the number of mistakes it made on the test dataset was 424.

2. d. The words that appear in more than 30 emails in the training dataset are:
['url', 'onc', 'dear', 'thi', 'on', 'nbsp', 'number', 'i', 'exmh', 'the', 'us', 'at', 'a', 'hello', 'httpaddr', 'from', 'nextpart', 'hi', 'in', 'begin']





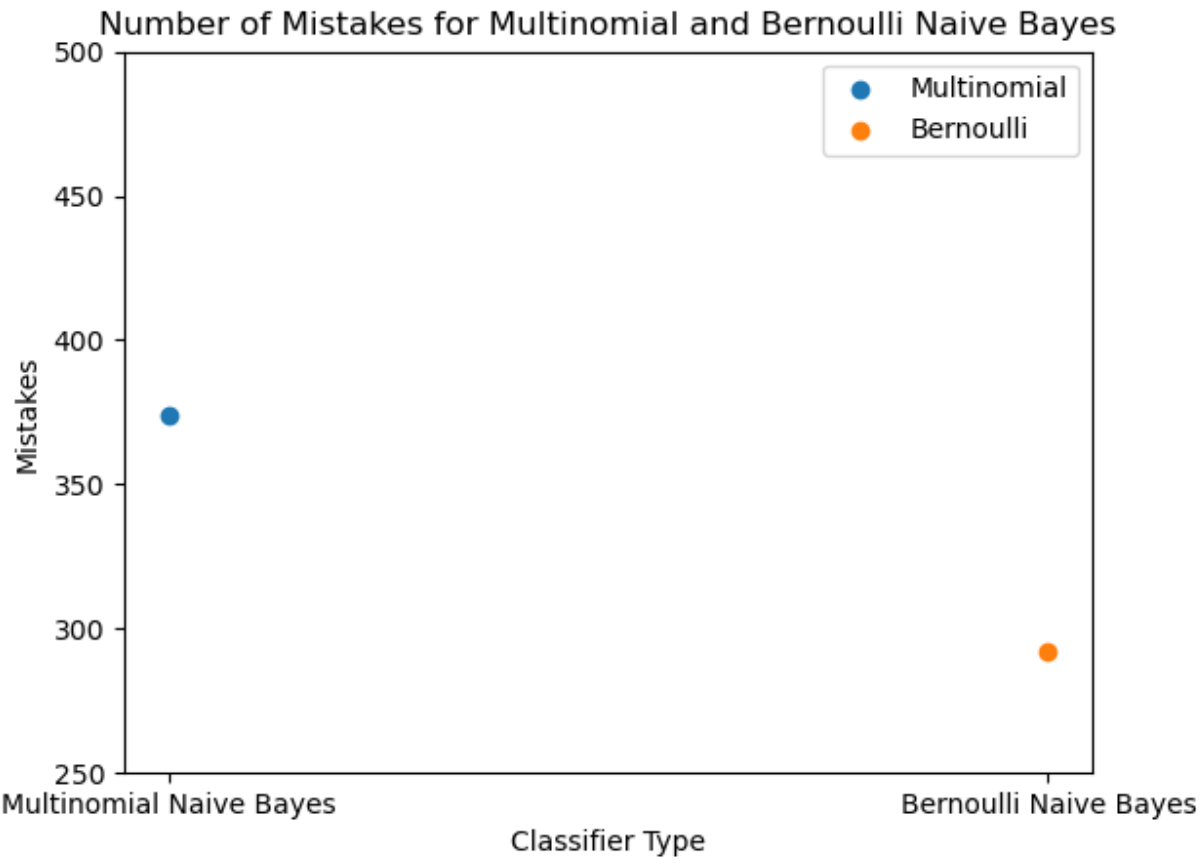
For the binary dataset, the 15 words with the most positive weight were:
at, in, httpaddr, a, from, the, begin, number, dear, us, exmh, on, hello, thi, url.

The 15 words with the most negative weight for the binary dataset were:
hi, i, onc, nextpart, nbsp, url, thi, hello, on, exmh, us, dear, number, begin, the.

For the count dataset, the 15 words with the most positive weight were:
at, in, httpaddr, the, hello, number, nbsp, a, on, from, dear, thi, nextpart, begin, us.

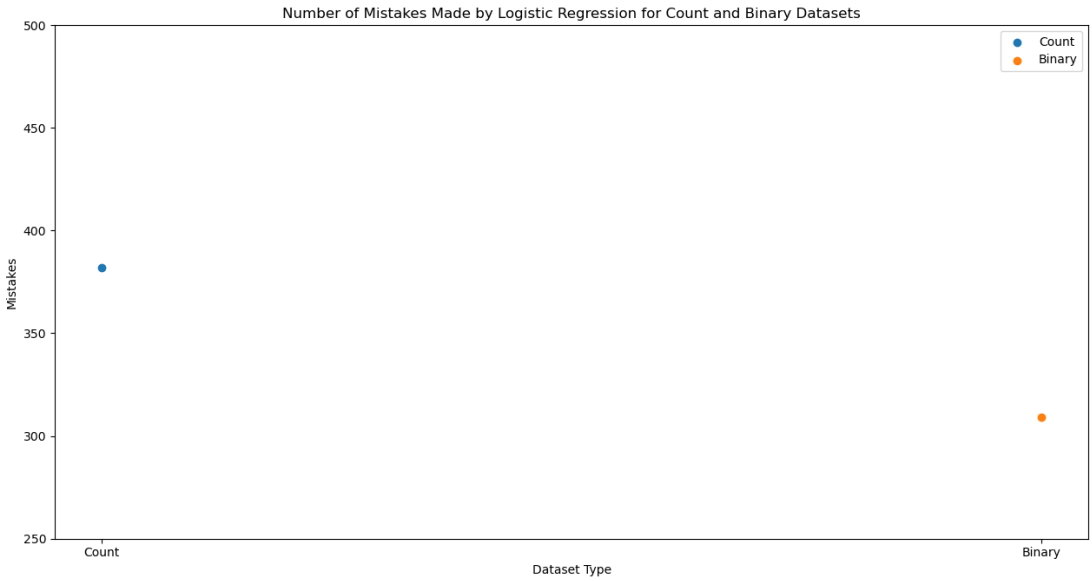
The 15 words with the most negative weight for the count dataset were:
hi, url, onc, exmh, i, us, begin, nextpart, thi, dear, from, on, a, nbsp, number.

3. a.



The Multinomial Naive Bayes sklearn model, which was used for the count dataset, had 374 errors on the test dataset.
The Bernoulli Naive Bayes sklearn model, which was used for the binary dataset, had 292 errors on the test dataset.

3b.



The logistic regression model made 382 mistakes on the count dataset, and 309 mistakes on the binary dataset.