3.

```
python BigramTester.py -f small_model.txt -t data/kafka.txt
Read 24944 words. Estimated entropy: 13.46
```

```
python BigramTester.py -f kafka_model.txt -t data/small.txt
Read 19 words. Estimated entropy: 10.29
```

(c)

```
❯ python .\BigramTester.py -f .\guardian_model.txt -t .\data\guardian_test.txt
Read 871878 words. Estimated entropy: 6.62
```

```
❯ python .\BigramTester.py -f .\austen_model.txt -t .\data\austen_test.txt
Read 10738 words. Estimated entropy: 6.97
```

4. (a) Batch GD:

```
Converged after 63 iterations
Model parameters:
0: -3.5565  1: 5.8156  2: -3.0479
                      Real class
                        0        1
Predicted class:  0 83831.000 3242.000
                  1  879.000 12046.000
```

- Accuracy = (83831 + 12046) / (83831 + 3242 + 879 + 12046) = 0.9588
- Precision (class 0) = 83831 / (83831 + 3242) = 0.9628
- Recall (class 0) = 83831 / (83831 + 879) = 0.9896
- Precision (class 1) = 12046 / (12046 + 879) = 0.9320
- Recall (class 1) = 12046 / (12046 + 3242) = 0.7879

(b) Mini-batch GD:

```
Model parameters:
0: -3.6068  1: 7.9998  2: -3.5832
                      Real class
                        0        1
Predicted class:  0 80674.000 2015.000
                  1 4036.000 13273.000
```

- Accuracy = (80674 + 13273) / (80674 + 2015 + 4036 + 13273) = 0.9395
- Precision (class 0) = 80674 / (80674 + 2015) = 0.9756
- Recall (class 0) = 80674 / (80674 + 4036) = 0.9524
- Precision (class 1) = 13273 / (13273 + 4036) = 0.8682
- Recall (class 1) = 13273 / (13273 + 2015) = 0.8683

(c) Stochastic GD:

```
Model parameters:
0: -6.8243  1: 12.5496  2: -31.1023
                      Real class
                        0        1
Predicted class:  0 83831.000 3242.000
                  1  879.000 12046.000
```

(d) Features: capitalized first letter, first token in sentence, length of the word

```
Converged after 62 iterations
Model parameters:
0: -1.9446  1: 5.8225  2: -3.0500  3: -1.6167
```

```
                      Real class
                        0        1
Predicted class:  0 83831.000 3242.000
                  1  879.000 12046.000
```

(e) Features: first token in sentence, number of capital letters, number of digits:

```
Model parameters:
0: -4.1558  1: -3.1467  2: 6.4065  3: 3.5969
                      Real class
                        0        1
Predicted class:  0 83604.000 2764.000
                  1 1106.000 12524.000
```

- Accuracy = (83604 + 12524) / (83604 + 2764 + 1106 + 12524) = 0.9613
- Precision (class 1) = 12524 / (12524 + 1106) = 0.9189
- Recall (class 1) = 12524 / (12524 + 2764) = 0.8192