Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

CS 440: NAÏVE BAYES CLASSIFICATION

PART 1: DIGIT CLASSIFICATION

1.1 Single Pixels as Features

Our implementation is quite simple. We created an object that would hold matrices for the binary 0 (white) and 1 (black) for each class. These matrices would be the size of an image. This meant that when we encountered a feature, we could simply find that exact location in a specific matrix and increment the count. For example, if we were looking at pixel (1,1) in a training image that represented the digit class 5 and the pixel was black, we would go to the black matrix for class 5 and increment its count. We followed this procedure for every feature of every training image. After doing this, we compared the testing data against the populated matrices for every single class, using the method described in the Assignment handout. For each testing image, we calculated its posterior probability for each class. We labelled it using the highest posterior probability. Our results are shown below.

We chose to use a smoothing factor of 1. This means that we assume that we have seen every value once more than its actual occurrence value. This made our results more accurate than using no smoothing factor. We chose not to use a larger smoothing factor; we did not want to over-count occurrences by more than 1. Using a smoothing factor of 1 means that if a particular feature never occurred for a given class, we could still operate under the assumption that it did occur once. This allows uncommon features to still be represented. Using a larger smoothing factor would have overrepresented features that do not occur very often. We did, however, build functionality to play around with smoothing factor. We made it possible to vary smoothing factor very easily. In our implementation, smoothing factor can be specified, which allows us to test different smoothing factors very easily. We found the best results with a smoothing factor of 1, as per our hypothesis. We want to represent all features, even if they occur uncommonly, but we do not want to over-represent uncommon features, which larger smoothing factors would do.

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

Results, as referenced above

Our overall accuracy rate for the digit data set is 77.1%

Below are the classification rates for each digit. The format is as follows:

(digit class, classification rate – truncated after 3 decimal points)

(0, 0.844)
(1, 0.962)
(2, 0.776)
(3, 0.790)
(4, 0.766)
(5, 0.673)
(6, 0.758)
(7, 0.726)
(8, 0.601)
(9, 0.800)

Below is the confusion matrix. Percentages are truncated after 2 decimal points.

	0:	1:	2:	3:	4:	5:	6:	7:	8:	9:
0:	84.44%	0.00%	1.11%	0.00%	1.11%	5.56%	3.33%	0.00%	4.44%	0.00%
1:	0.00%	96.30%	0.93%	0.00%	0.00%	1.85%	0.93%	0.00%	0.00%	0.00%
2:	0.97%	2.91%	77.67%	3.88%	0.97%	0.00%	5.83%	0.97%	4.85%	1.94%
3:	0.00%	2.00%	0.00%	79.00%	0.00%	3.00%	2.00%	6.00%	2.00%	6.00%
4:	0.00%	0.93%	0.00%	0.00%	76.64%	0.00%	2.80%	0.93%	1.87%	16.82%
5:	2.17%	2.17%	1.09%	13.04%	3.26%	67.39%	1.09%	1.09%	2.17%	6.52%
6:	1.10%	6.59%	4.40%	0.00%	4.40%	5.49%	75.82%	0.00%	2.20%	0.00%
7:	0.00%	5.66%	2.83%	0.00%	2.83%	0.00%	0.00%	72.64%	2.83%	13.21%
8:	1.94%	0.97%	2.91%	13.59%	1.94%	5.83%	0.00%	0.97%	60.19%	11.65%
9:	1.00%	1.00%	1.00%	3.00%	9.00%	2.00%	0.00%	2.00%	1.00%	80.00%

Below are the test examples with the highest and lowest posterior probabilities for each digit class. White regions are represented as zeros and colored regions are represented as ones. See below.

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:    0
High Frequency: -146.323892153

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000011110000000000
0000000000001111111000000000
0000000000011111111110000000
0000000000111111111111000000
0000000001111111101111000000
0000000011111111001111100000
0000000011111000000111100000
0000000111110000000011100000
0000000111110000000011100000
0000000111100000000011100000
0000000111100000000011100000
0000000111100000000011100000
0000000111100000000111100000
0000000111100000001111000000
0000000111110000011111000000
0000000111111000111111000000
0000000011111111111110000000
0000000011111111111100000000
0000000001111111111000000000
0000000000011111100000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   0
Low Frequency:  -552.720115536

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000111100000000000000000000
0000111100000000000000000000
0000111100000001111000000000
0000111100000011111000000000
0000111100000011111000000000
0000111110000011111100000000
0000111110000011111100000000
0000111110000011111100000000
0000011110000011111100000000
0000001111000011111110000000
0000001111111111111110000000
0000000111111111111110000000
0000000011111111111110000000
0000000001111111111110000000
0000000000011111111110000000
0000000000000000111111000000
0000000000000000111111110000
0000000000000000111111110000
0000000000000000011111110000
0000000000000000001111110000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   1
High Frequency: -71.6348012014

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000011100000000000
0000000000000011110000000000
0000000000000011110000000000
0000000000000011110000000000
0000000000000011110000000000
0000000000000011100000000000
0000000000000111100000000000
0000000000000111100000000000
0000000000000111100000000000
0000000000000111100000000000
0000000000000111000000000000
0000000000001111000000000000
0000000000001111000000000000
0000000000001111000000000000
0000000000011110000000000000
0000000000011110000000000000
0000000000111110000000000000
0000000000111110000000000000
0000000000111110000000000000
0000000000011100000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   1
Low Frequency:  -902.826929454

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000011111111110000000000
0000001111111111111100000000
0000001111111111111111000000
0000011111111111111111100000
0000011111111111111111110000
0000001100000000111111110000
0000000000000000001111110000
0000000000000000000111110000
0000000000000000000111110000
0000001111111000000111110000
0000011111111100001111110000
0000011111111111011111110000
0000111111111111111111110000
0000111111111111111111100000
0000111110011111111111100000
0000111111111111111111000000
0000111111111111111111000000
0000011111111111111000000000
0000000111111111000000000000
0000000011111100000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   2
High Frequency: -169.157830278

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000111111000000000000
0000000000111111110000000000
0000000001111111111000000000
0000000001111001111100000000
0000000001100000111100000000
0000000000000000011100000000
0000000000000000011100000000
0000000000000000001100000000
0000000000000000001100000000
0000000000000000001100000000
0000000000000000011100000000
0000000000000000011100000000
0000000000011111011100000000
0000000001111111111000000000
0000000011111111111000000000
0000000011111111111110000000
0000000011100111111111110000
0000000011111111100111100000
0000000011111111000000000000
0000000001111100000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   2
Low Frequency:  -482.154509687

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000011111100
0000000000000000000011111100
0000000011111111111111111100
0000000111111111111111111100
0000001111111111111111110000
0000001111111111111110000000
0000001111111111110000000000
0000001111111111100000000000
0000000111111111110000000000
0000000000000111110000000000
0000000000000011110000000000
0000000000000011110000000000
0000000000000011110000000000
0000000000000001110000000000
0000000000000011110000000000
0000000000000111111000000000
0000000011111111111000000000
0000000011111111110000000000
0000000011111111110000000000
0000000001111111000000000000
0000000000000000000000000000
0000000000000000000000000000
```

```
Category:   3
High Frequency: -145.279996263

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000001111111100000000000
0000000001111111111000000000
0000000000001111111100000000
0000000000000000111100000000
0000000000000000011110000000
0000000000000000111100000000
0000000000000011111000000000
0000000000000111110000000000
0000000000001111100000000000
0000000000011111110000000000
0000000000011111111000000000
0000000000011001111000000000
0000000000000000111000000000
0000000000000000111000000000
0000000000000000111000000000
0000000000000001111000000000
0000000000000011111000000000
0000011111111111110000000000
0000011111111111100000000000
0000001111111111000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   3
Low Frequency:  -547.318118526

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000011111110000000
0000000000111111111111000000
0000000011111111111111100000
0000011111111111111111100000
0000111111111111110111100000
0001111110000000000011100000
0001111100000000000011100000
0001111000000000000111100000
0001111000000001111111100000
0001111111111111111111100000
0001111111111111111111100000
0000111111111111100111100000
0000001111111100000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   4
High Frequency: -139.655360285

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000010000000110000000
0000000000111000000110000000
0000000000111000000110000000
0000000001111000011110000000
0000000011110000011110000000
0000000011110000011110000000
0000000011100000011110000000
0000000011100000111110000000
0000000011000011111110000000
0000001111111111111000000000
0000000111111111111000000000
0000000000111111110000000000
0000000000000011100000000000
0000000000000011100000000000
0000000000000011100000000000
0000000000000111100000000000
0000000000000111000000000000
0000000000000111000000000000
0000000000000111000000000000
0000000000000111000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   4
Low Frequency:  -511.213498046

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000111111111000000000
0000000001111111111100000000
0000000011111111111111000000
0000001111111111111111110000
0000001111111110000011110000
0000111111100000000011110000
0000111111000000000011110000
0000111110000000000011110000
0000111100000000000011110000
0000111100000000000000110000
0000111100000000000000110000
0000111100000000000000110000
0000111100000000000000110000
0000111100000000000000110000
0000011111000000000011110000
0000001111110000000011110000
0000001111111111111111110000
0000001111111111111111000000
0000000000011111111100000000
0000000000001111110000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   5
High Frequency: -163.633260013

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000111100000
0000000000001111111111100000
0000000000001111111111000000
0000000000001111111100000000
0000000000011110000000000000
0000000000011100000000000000
0000000000111000000000000000
0000000000111111000000000000
0000000001111111100000000000
0000000001111111100000000000
0000000000100001110000000000
0000000000000001110000000000
0000000000000001110000000000
0000000000000000110000000000
0000000111000000111000000000
0000000111000001110000000000
0000000111000001110000000000
0000000111100011110000000000
0000000011111111100000000000
0000000000011111000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   5
Low Frequency:  -489.904511857

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000011111110000000
0000000000111111111111000000
0000000011111111111111100000
0000011111111111111111100000
0000111111111111110111100000
0001111110000000000011100000
0001111100000000000011100000
0001111000000000001111100000
0001111000000001111111100000
0001111111111111111111100000
0001111111111111111111100000
0000111111111111100111100000
0000001111111100000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000111100000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   6
High Frequency: -120.651413332

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000111000000000
0000000000000011111000000000
0000000000000111110000000000
0000000000001111100000000000
0000000000001111100000000000
0000000000011111000000000000
0000000000111110000000000000
0000000000111100000000000000
0000000000111100000000000000
0000000001111100000000000000
0000000001111000000000000000
0000000001111000111110000000
0000000011111001111110000000
0000000011111111111110000000
0000000011110111111110000000
0000000011111111111110000000
0000000011111111111000000000
0000000001111111110000000000
0000000000111111100000000000
0000000000001111000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   6
Low Frequency:  -619.121045481

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000111111111100000000
0000000001111111111110000000
0000000011111111111111000000
0000001111111111111111110000
0000001111111110000011110000
0000111111100000000011110000
0000111111000000000011110000
0000111110000000000011110000
0000111100000000000011110000
0000111100000000000001100000
0000111100000000000001100000
0000111100000000000001100000
0000111100000000000001100000
0000111100000000000001100000
0000011111000000000011110000
0000001111110000000011110000
0000001111111111111111110000
0000001111111111111111000000
0000000000011111111100000000
0000000000001111110000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   7
High Frequency: -116.163705357

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000011110111111000000000
0000000011111111111100000000
0000000011111111111100000000
0000000000111000011100000000
0000000000000000111100000000
0000000000000000111100000000
0000000000000001111000000000
0000000000000001111000000000
0000000000000001111000000000
0000000000000011110000000000
0000000000000011110000000000
0000000000000011100000000000
0000000000000111100000000000
0000000000000111000000000000
0000000000001111000000000000
0000000000001110000000000000
0000000000011100000000000000
0000000000011100000000000000
0000000000011100000000000000
0000000000011100000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   7
Low Frequency:  -632.810357967

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000011111111110000000000
0000001111111111111100000000
0000001111111111111111000000
0000011111111111111111100000
0000011111111111111111110000
0000001110000000111111110000
0000000000000000001111110000
0000000000000000000111110000
0000000000000000000111110000
0000001111111000000111110000
0000011111111100001111110000
0000011111111111011111110000
0000111111111111111111110000
0000111111111111111111100000
0000111110011111111111100000
0000111111111111111111000000
0000111111111111111111000000
0000011111111111111000000000
0000000111111111000000000000
0000000011111100000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   8
High Frequency: -138.566348953

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000111100000000
0000000000000011111110000000
0000000000001111111110000000
0000000000011111111110000000
0000000000111111011110000000
0000000000111100011110000000
0000000000111000001111000000
0000000000111000011111000000
0000000000111111111111000000
0000000000111111111100000000
0000000000011111110000000000
0000000000111111000000000000
0000000001111111100000000000
0000000001111111100000000000
0000000011110011110000000000
0000000011110011110000000000
0000000011110111100000000000
0000000011111111100000000000
0000000001111111100000000000
0000000000111111000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

```
Category:   8
Low Frequency:  -525.03754002

0000000000000000000000000000
0000000000000000000000000000
0000000000111110000000000000
0000000011111110000000000000
0000000111111000000000000000
0000000111000000000000000000
0000001110000000000000000000
0000001110000000000000000000
0000001110000000000000000000
0000001110000000000000000000
0000001110000000000000000000
0000001110000000000000000000
0000000110000000000000000000
0000000111000000000111000000
0000000111000000011111110000
0000000011100000111111111000
0000000001110000111000111000
0000000001111001111000011100
0000000000111110111000011100
0000000000001111111100111000
0000000000000011111111111000
0000000000000000011111110000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

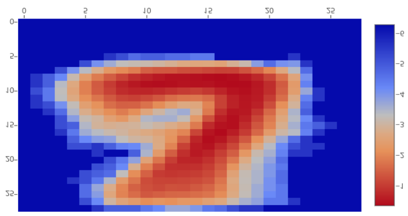
```
Category:   9
High Frequency: -125.113036446

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000001111110000000000
0000000000011111111000000000
0000000001111111111100000000
0000000011111000111100000000
0000000011110001111100000000
0000000111100001111100000000
0000000111100001111100000000
0000000111100111111000000000
0000000011111111111000000000
0000000001111111110000000000
0000000000111111110000000000
0000000000000111100000000000
0000000000000111100000000000
0000000000000111000000000000
0000000000001111000000000000
0000000000001111000000000000
0000000000001110000000000000
0000000000001110000000000000
0000000000001110000000000000
0000000000001110000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

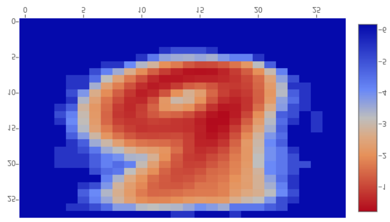
```
Category:   9
Low Frequency:  -649.059311148

0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000001111111111000000000
0000000001111111111111000000
0000000011110000111111000000
0000000111100000000111110000
0000000111000000000011110000
0000001111000000000001110000
0000001110000000000000111000
0000011110000000000000111000
0000011100000000000000111000
0000011100000000000000111000
0000011100000000000000111000
0000011100000000000001111000
0000011100000000000001111000
0000011100000000000111110000
0000011100000000001111110000
0000011100000000111111100000
0000011110000001111111000000
0000001111111111111100000000
0000001111111111111000000000
0000000111111111000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
0000000000000000000000000000
```

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

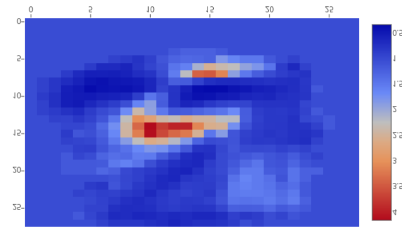
We also looked at odds ratios for the following sets of numbers: [(7,9),(4,9),(5,3),(8,3)]. The results are below, shown as heat maps of the log likelihoods.



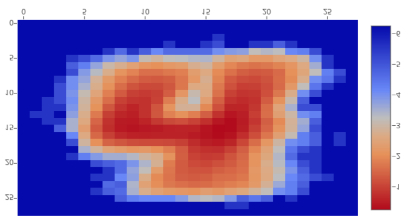
7



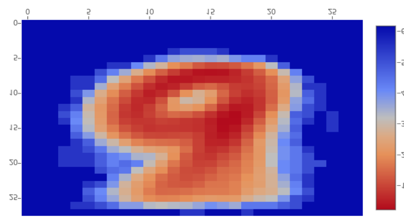
9



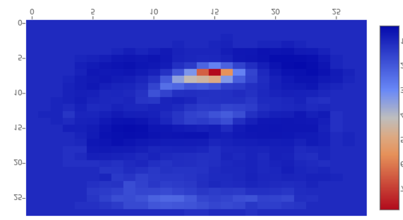
7 over 9



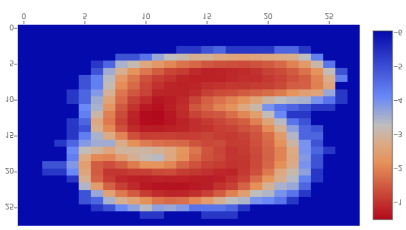
4



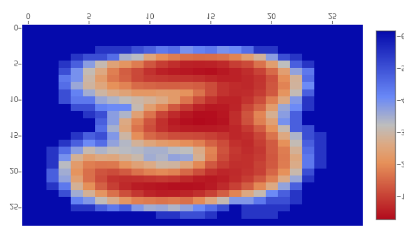
9



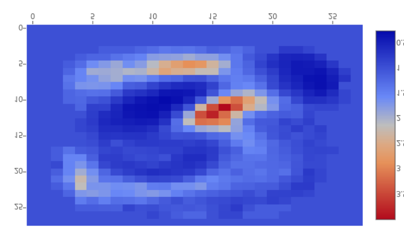
4 over 9



5

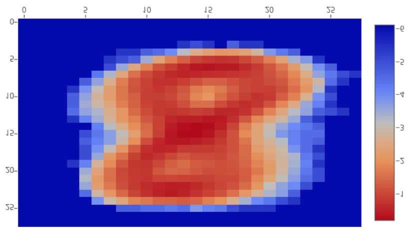


3

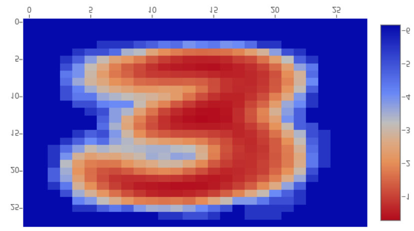


5 over 3

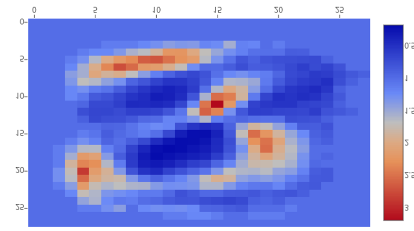
Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)



8



3



8 over 3

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

Extra Credit for Part 1

We experimented with using ternary features. This is for bonus points. Our implementation was similar, but we had matrices to represent white, gray, and black rather than just white and black. Our results reflected a small improvement.

Our overall accuracy improved to 77.6%.

Below are the classification rates for each digit. The format is as follows:

(digit class, classification rate – truncated after 3 decimal points)

(0, 0.833)
(1, 0.953)
(2, 0.766)
(3, 0.800)
(4, 0.775)
(5, 0.684)
(6, 0.780)
(7, 0.735)
(8, 0.621)
(9, 0.800)

The confusion matrix is shown below.

	0:	1:	2:	3:	4:	5:	6:	7:	8:	9:
0:	83.33%	0.00%	1.11%	0.00%	0.00%	6.67%	4.44%	0.00%	4.44%	0.00%
1:	0.00%	95.37%	0.00%	0.00%	0.00%	1.85%	0.93%	0.00%	1.85%	0.00%
2:	0.97%	2.91%	76.70%	3.88%	0.97%	0.97%	5.83%	1.94%	4.85%	0.97%
3:	0.00%	2.00%	0.00%	80.00%	0.00%	3.00%	2.00%	5.00%	3.00%	5.00%
4:	0.00%	0.00%	0.00%	0.00%	77.57%	0.93%	1.87%	0.93%	1.87%	16.82%
5:	2.17%	1.09%	1.09%	13.04%	3.26%	68.48%	1.09%	1.09%	2.17%	6.52%
6:	0.00%	4.40%	4.40%	0.00%	4.40%	5.49%	78.02%	0.00%	3.30%	0.00%
7:	0.00%	5.66%	2.83%	0.00%	2.83%	0.00%	0.00%	73.58%	2.83%	12.26%
8:	0.97%	1.94%	2.91%	11.65%	1.94%	8.74%	0.00%	0.97%	62.14%	8.74%
9:	1.00%	1.00%	0.00%	2.00%	10.00%	2.00%	0.00%	2.00%	2.00%	80.00%

It appears that ternary features did not make a significant improvement upon classification. However, we do believe that weighting gray values differently could be worth further exploration.

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

We applied our Naïve Bayes Classifier to the face data set. The implementation was very similar to the digit data set, except that there were only two classes and that there were more features per image. This is for bonus points.

Our overall accuracy was 90.6%.

Below is the classification rate for each class. The format is as follows, with 0 being not face and 1 being a face:

(face class, classification rate – truncated after 3 decimal points)

(0, 0.883)
(1, 0.931)

The confusion matrix is shown below.

	0:	1:
0:	88.31%	11.69%
1:	6.85%	93.15%

This whole section was for bonus points.

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

PART 2: DOCUMENT CLASSIFICATION

2.1: For Three-Unit Students

Results

SPAM DATA SET

Multinomial

Accuracy 97.30769%

Confusion matrix is below.

Email/Spam – Multinomial

	0	1
0:	96.15%	3.85%
1:	1.54%	98.46%

Top 20 Words for Normal Email

language
university
s
linguistic
de
information
conference
workshop
email
paper
e
english
one
please
include
edu
http
research
abstract
address

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

Top 20 Words for Spam Email

email
s
order
report
our
address
mail
program
send
free
money
list
receive
name
business
one
d
work
com
nt

Bernoulli

Accuracy 96.53846%

Confusion matrix is below.

Email/Spam – Bernoulli

	0	1
0:	93.85%	6.15%
1:	0.77%	99.23%

Top 20 Words for Normal Email

language
university
s
information
linguistic
http
email
please
e
follow
fax
include
one
english
call
research
www
word
address
interest

Top 20 Words for Spam Email

our
s
free
please
email
mail
one
address
list
com
receive
http
us
send
day
information
remove
here
over
want

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

MOVIE DATA SET

Multinomial

Accuracy 75.70000%

Confusion matrix is below.

Movie – Multinomial

	-1	1
-1:	75.00%	25.00%
1:	23.60%	76.40%

Top 20 Words for Negative Reviews

movie
film
like
one
--
bad
story
much
time
even
characters
good
little
would
comedy
never
nothing
makes
plot
make

Top 20 Words for Positive Reviews

film
movie
--
one
like
story
good
comedy
way
even
time
best
much
performances
funny
make
life
us
makes
characters

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

Bernoulli

Accuracy 75.70000%

Confusion matrix is below.

Movie – Bernoulli

	-1	1
-1:	72.00%	28.00%
1:	20.60%	79.40%

Top 20 Words for Negative Reviews

movie
film
like
one
story
much
--
bad
time
even
characters
little
good
would
comedy
nothing
makes
plot
never
make

Top 20 Words for Positive Reviews

film
movie
one
like
--
story
comedy
way
even
good
best
time
much
performances
funny
makes
life
make
characters
work

Part 2.2: For Four-Unit Students (Extra Credit for Us)

Though we are three-unit students, we attempted the four-unit problem. This is for bonus points.

Results

8 CATEGORY DATA SET

Multinomial

Accuracy 92.01521%

Confusion matrix is below.

8 Category – Multinomial

	0	1	2	3	4	5	6	7
0:	97.06%	0.00%	0.00%	0.00%	2.94%	0.00%	0.00%	0.00%
1:	0.00%	96.97%	0.00%	0.00%	0.00%	0.00%	0.00%	3.03%
2:	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
3:	0.00%	3.57%	3.57%	82.14%	0.00%	3.57%	0.00%	7.14%
4:	2.13%	2.13%	2.13%	0.00%	93.62%	0.00%	0.00%	0.00%
5:	0.00%	20.00%	0.00%	0.00%	0.00%	80.00%	0.00%	0.00%
6:	0.00%	0.00%	2.17%	0.00%	0.00%	2.17%	95.65%	0.00%
7:	0.00%	13.79%	0.00%	3.45%	0.00%	6.90%	0.00%	75.86%

Top 20 Words for sci.space

space
nt
would
one
launch
nasa
earth
subject
like
us
system
also
writes
could
time
data
first
orbit
edu
mission

Top 20 Words for comp.sys.ibm.pc.hardware

drive
scsi
nt

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

ide
one
card
drives
controller
system
disk
subject
use
would
edu
hard
bus
get
m
data
also

Top 20 Words for rec.sport.baseball

nt
would
year
edu
writes
one
game
good
team
subject
last
article
think
players
like
baseball
games
better
well
time

Top 20 Words for comp.windows.x

x
window
use
nt
subject
file
server
also
available
get
edu
motif
version
system
program
sun
c
one
m
windows

Top 20 Words for talk.politics.misc

nt
would
people
q
one
mr
think
writes
president
article
government
stephanopoulos
know
us
edu
like
subject
going
right
get

Top 20 Words for misc.forsale

new
edu
dos
sale
appears
art
subject
wolverine
shipping
cover
price
one
list
comics
drive
nt
hulk
good
vs
system

Top 20 Words for rec.sport.hockey

nt
game
team
hockey
would
play
subject
period
season
nhl
games
one
first
year
think
players
get
la
edu
like

Top 20 Words for comp.graphics

image
jpeg
edu
nt
file
images
data
also
graphics
software
available
use
one
program
files
format
get
version
system
ftp

Bernoulli

Accuracy 84.41065%

Confusion matrix is below.

8 Category – Bernoulli

	0	1	2	3	4	5	6	7
0:	82.35%	0.00%	2.94%	0.00%	14.71%	0.00%	0.00%	0.00%
1:	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
2:	0.00%	0.00%	100.00%	0.00%	0.00%	0.00%	0.00%	0.00%
3:	0.00%	25.00%	3.57%	64.29%	0.00%	0.00%	0.00%	7.14%
4:	0.00%	0.00%	2.13%	0.00%	97.87%	0.00%	0.00%	0.00%
5:	0.00%	40.00%	0.00%	0.00%	10.00%	50.00%	0.00%	0.00%
6:	0.00%	2.17%	2.17%	0.00%	0.00%	0.00%	95.65%	0.00%
7:	3.45%	48.28%	0.00%	0.00%	6.90%	0.00%	0.00%	41.38%

Top 20 Words for sci.space

subject
would
nt
space
writes
article
one
like
could
also
get
think
time
us
new
much
way
m
edu
see

Top 20 Words for comp.sys.ibm.pc.hardware

subject
nt
one
would
writes
use
get
know
article
card
also
like
m
two
system
edu
drive
work
problem
could

Top 20 Words for rec.sport.baseball

subject
nt
writes
article
edu
would
one
last
year
like
baseball
good
think
get
time

m	game	first
know	team	

Top 20 Words for comp.windows.x

subject	article	know
x	using	code
nt	one	m
use	like	set
writes	would	email
get	also	help
window	problem	

Top 20 Words for talk.politics.misc

subject	like	government
nt	edu	could
writes	us	know
article	even	make
people	think	time
would	m	much
one	get	

Top 20 Words for misc.forsale

subject	price	want
sale	nt	used
edu	one	good
new	get	use
shipping	condition	sell
please	like	etc
email	list	

Top 20 Words for rec.sport.hockey

subject	game	would
nt	hockey	one
team	writes	article

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

like
play
first
think
go
get
nhl
year
games
time
last

Top 20 Words for comp.graphics

subject
nt
one
would
writes
also
like
article
graphics
edu
use
get
know
computer
need
could
think
program
m
two

Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

Extra Credit for Part 2

We did word clouds for every single category for ever dataset. This is for bonus points. They are shown below.

Normal Email

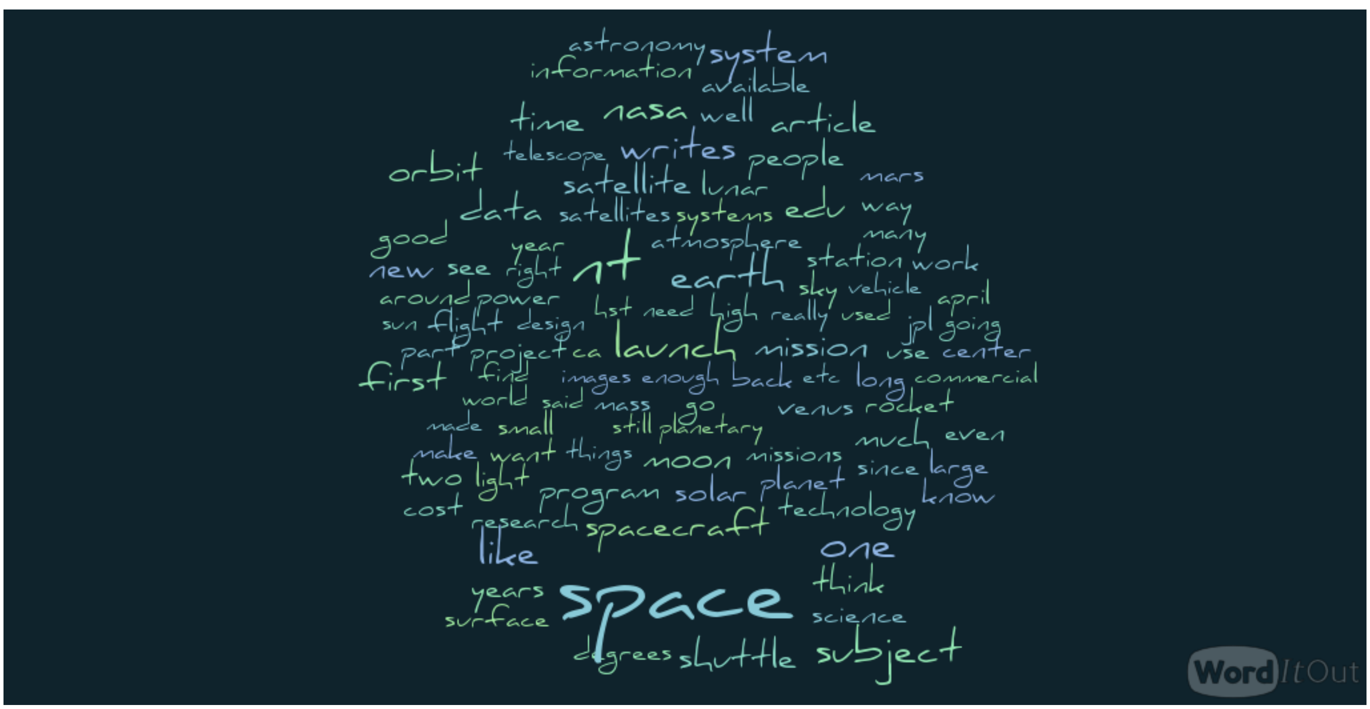
Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

Spam Email

Negative Movie Review

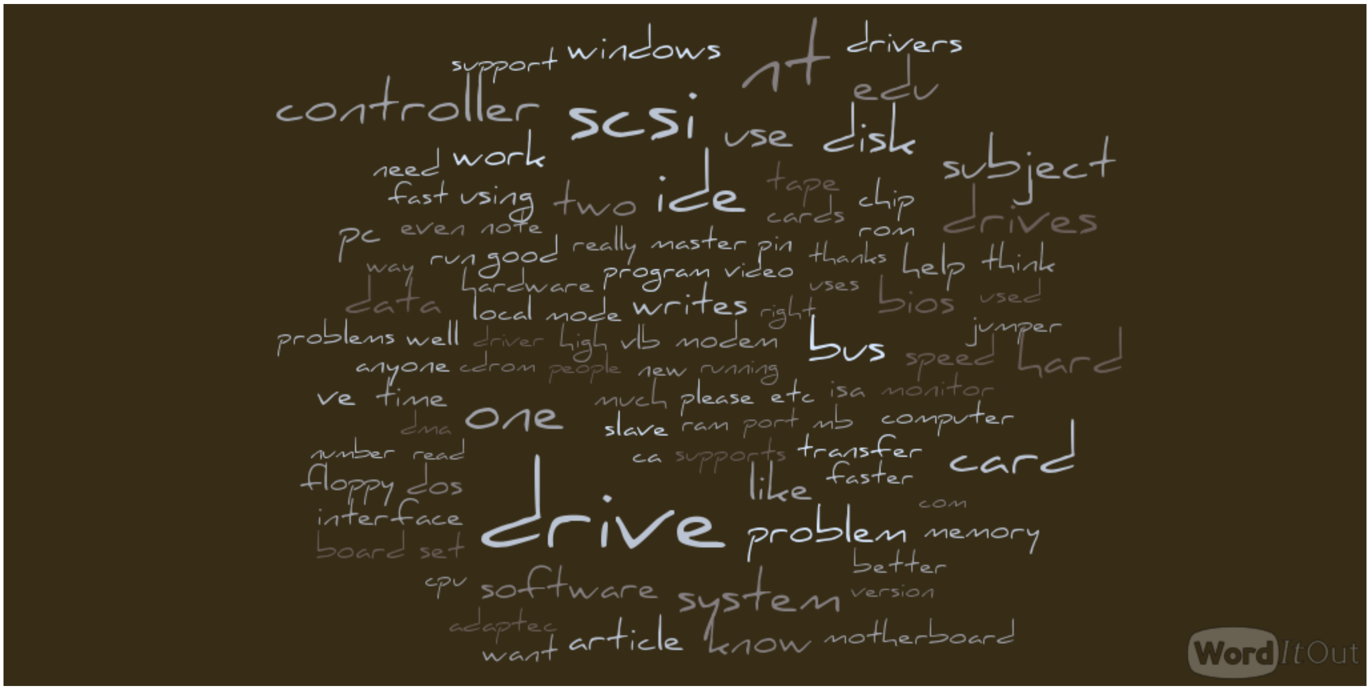
Positive Movie Reiview

sci.space

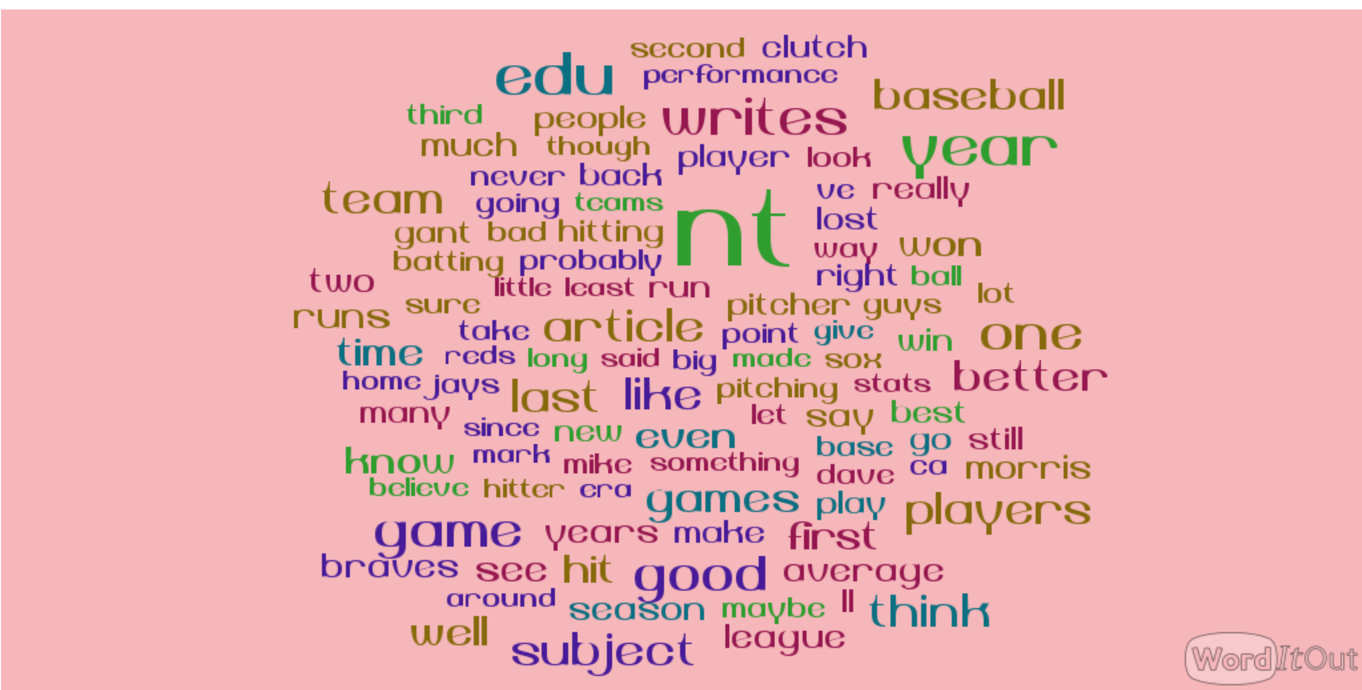


Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

comp.sys.ibm.pc.hardware

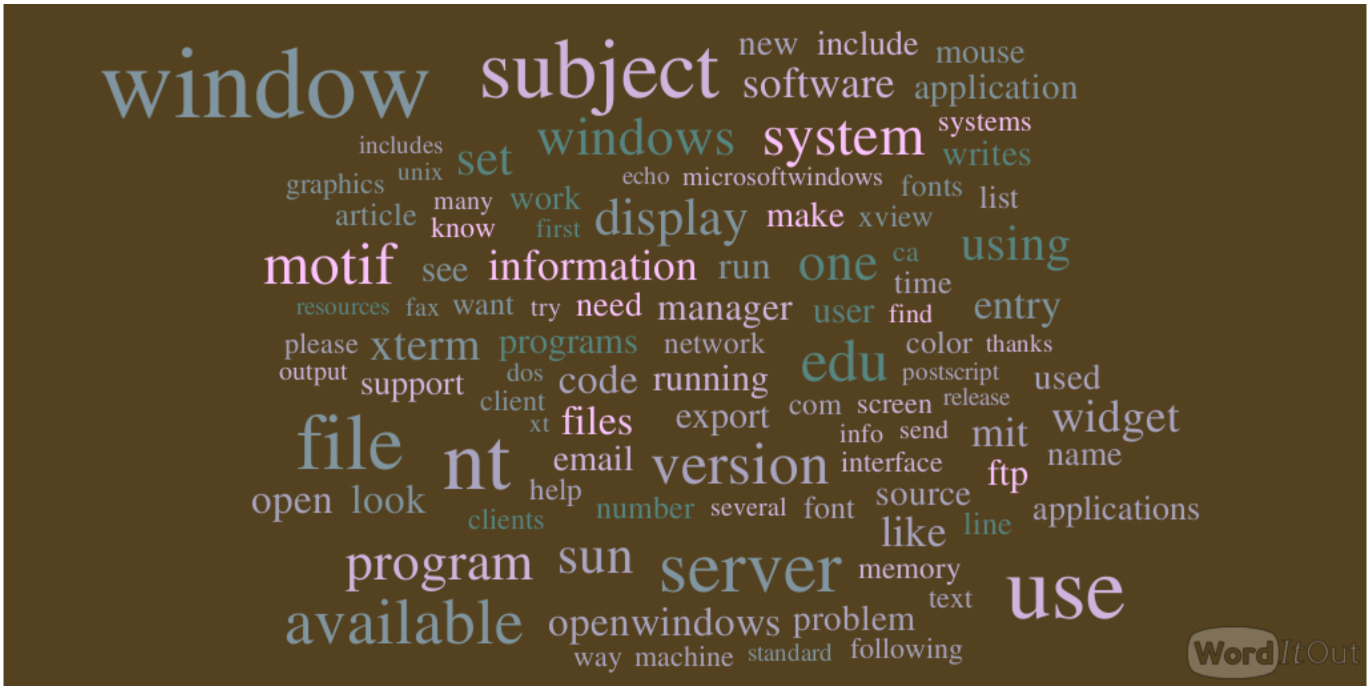


rec.sport.baseball

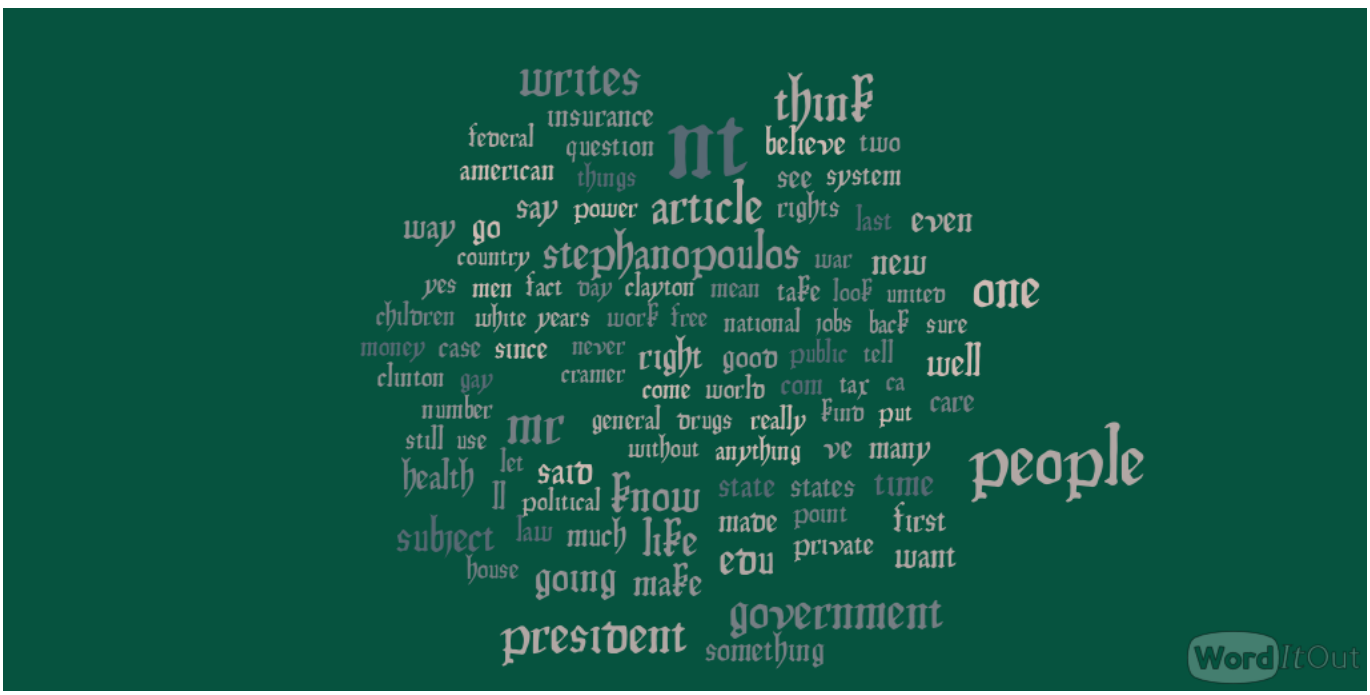


Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

comp.windows.x

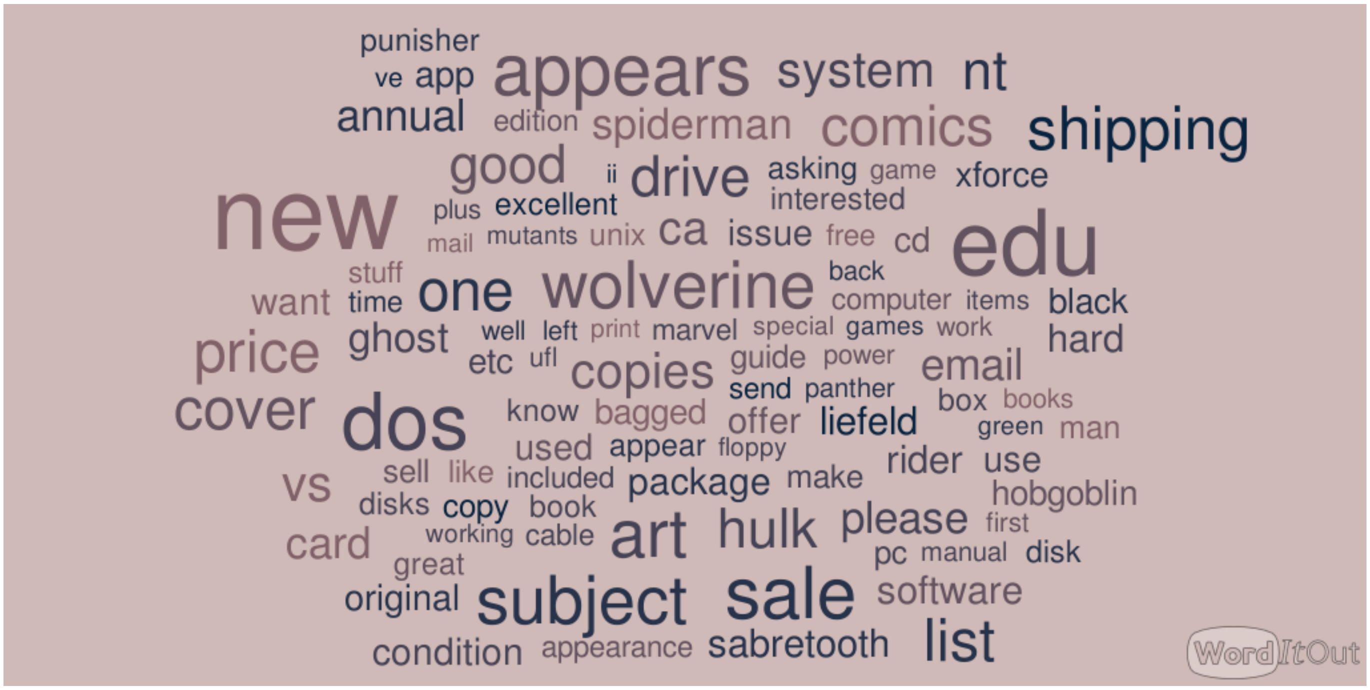


talk.politics.misc

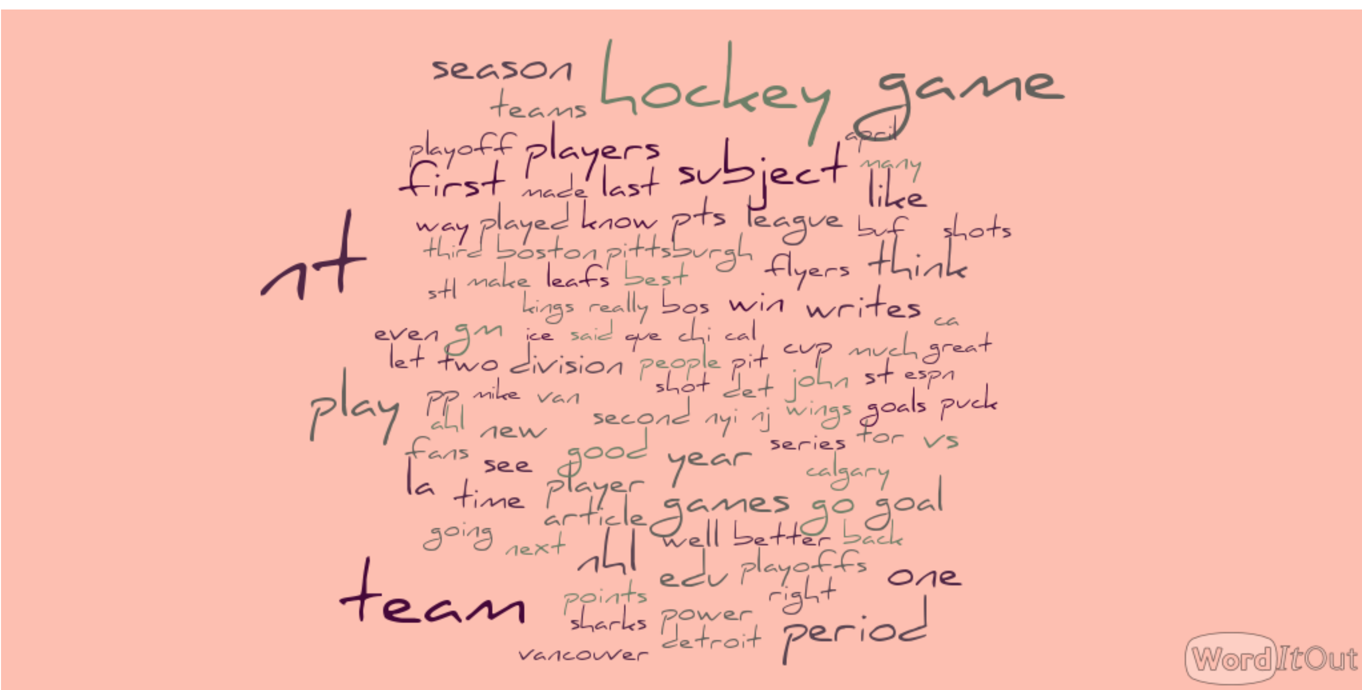


Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

misc.forsale

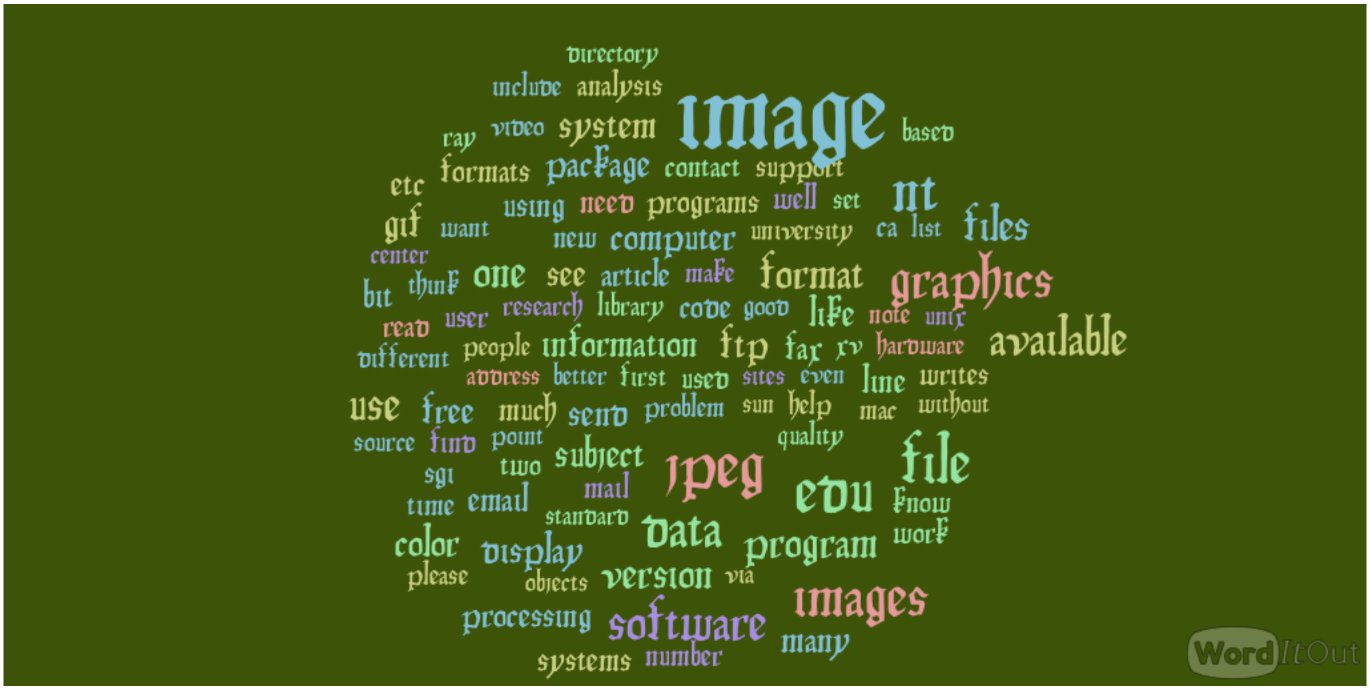


rec.sport.hockey



Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

comp.graphics



Jakub Klapacz – jklapac2 (3 units)
Abhishek Nigam – adnigam2 (3 units)

INDIVIDUAL CONTRIBUTIONS

Abhishek Nigam

I setup the infrastructure for part 1 and developed the training model for part 2. I generated the Top 20 Words list. We worked on the rest together.

Jakub Klapacz

I setup the infrastructure for part 2 and developed the training model for part 1. I generated the confusion matrices. We worked on the rest together.