Minhash Signatures

Andey Robins Tuttle

1. Execution Instructions

For more complete execution information, see the `README.md` file in the root of the submitted zip file. Use the following command to build and run the example from class. Expected output is shown after the command.

```
$ go run main.go -class

Expected Matrix:
0 3 1
1 0 0

Calculated Matrix:
0 3 1
1 0 0

Matrices were equal. Everything worked!

The true Jaccard similarity is 0.2 while the estimated similarity is 0
```

And for execution using a randomized set of column vectors:

```
$ go run main.go -L 10007 -Q 1000
Running program with values L=10007, Q=1000

Generated sparse columns
Constructed sparse matrix
Constructed hashes
Approximated Jaccard = 0.0049482451013294895
True Jaccard         = 0.004997187665417682
```

2. Class Example

All computations performed manually in class were correctly mirrored in the code. Running the code with command `go run main.go -class` will invoke the class example, printing both the calculated sig matrix and estimated Jaccard Similarity. The output produced locally on my computer (Macbook Pro, macOS Monterey, 12.6, arm64) is also reproduced in §1 above.

3. Scaling Results

The following results were obtained by using an initial prime $L = 10007$ and varying the value of Q from 2 to 1000. Further, a set seed of `-seed = 22` was used to generate the graphics. The handy script used to generate the data is included as `run_many.sh`.

After each signature approximation matrix is calculated, the Jaccard Similarity for each pair of columns i, j in both the sparse matrix and the signature matrix are calculated. Then we store the difference between the average Jaccard Similarity between columns and the average estimated Jaccard Similarity. These are tracked to calculate the percentage error; this percent error is then presented in Figure 1.

Asymptotically, the percent error levels out and approaches the true Jaccard similarity with a ≈1% error. The percent error was calculated with the standard formula of $|\frac{x_a - x_e}{x_e}| * 100$ where x_a is the approximated similarity between two rows and x_e is the exact similarity between two rows.

Figure 2 shows the difference between the estimated similarity and the true similarity. Since the value asymptotically levels out slightly below a difference of 0, we see that the signature matrix minhash approach slightly under-estimates the similarity between two columns in these instances.

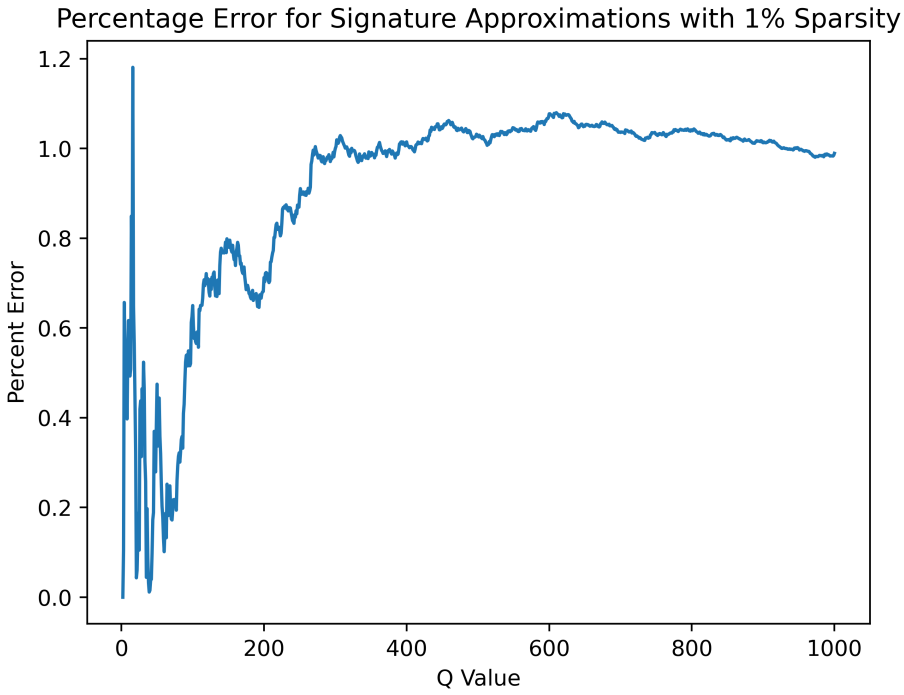

Figure 1: The percentage error calculated based upon the estimated Jaccard similarity and the actual Jaccard similarity for randomized column vectors containing approximately 1% non-zero data. The value for q is varied to visualize the asymptotic relationship between the estimated similarity and true similarity.

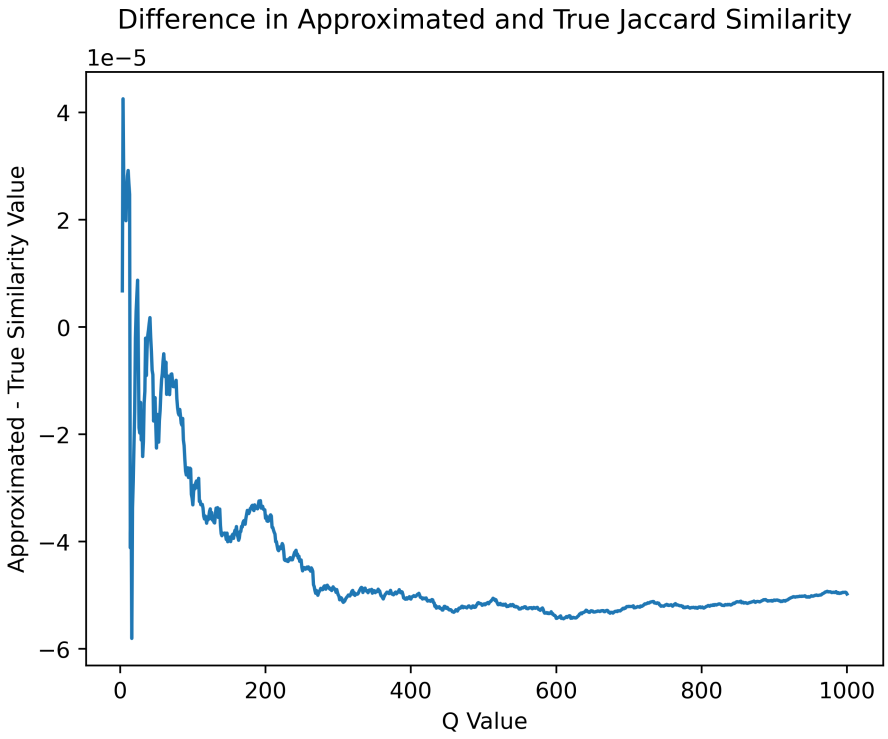

Figure 2: The difference between the estimated Jaccard similarity and the actual Jaccard similarity for randomized column vectors containing approximately 1% non-zero data. The value for q is varied to visualize the asymptotic relationship between the estimated similarity and true similarity.