Deduplicating Emails

Goal: To reduce the time and cost of scanning through emails

Task: Find any duplicate emails and remove them

Method

We computed the hash of all of the emails and grouped emails according to hash. This reduced 9,199 emails into 4,712 emails (49%) reduction.

We also tried looking at forwarded emails, though we find that most forwarded emails have added commentary that makes them unique.

Results

Among the 4,487 duplicated emails, they can grouped into 2,064 duplicated groups. Most of them are system duplicates, where everything is the same aside from email ID.

Pattern	Groups	%	What Differs
System duplicates	2,022	98.0%	Only email ID
Broadcast/format variations	31	1.5%	Recipient address format
Automated newsletters	5	0.2%	Subject date
Error messages/spam	5	0.2%	Everything except body

Note that the distribution of duplicated emails could be specific to each client, though it may be the case that another client could have smaller duplication rates.

Results for forwarded emails

After deduplication, we ran an analysis on forwarded emails.

Total forwarded emails:	3,040 (33% of dataset)

```
├─ Already caught as EXACT duplicates:      1,871 (61.5%)
│   └─ These are forwards sent multiple times unchanged
│
└─ Remaining after exact dedup:             1,169 (38.5%)
   │
   ├─ Caught by forwarding cleaning:          31 (2.7%)
   │   └─ Same person → multiple recipients
   │
   └─ Remain unique:                        1,138 (97.3%)
      ├─ Have added commentary:               877 (77.1%)
      ├─ Chained forwards:                    199 (17.5%)
      ├─ Original not in dataset:             955 (84.0%)
      └─ Other variations:                    62 (5.4%)
```

Forwarding cleaning (where we remove the forward headers and find exact matches) only identified 2.7% of emails remaining after deduplications. The rest (97.3%) usually have added commentary or the original email is not in the dataset (or both).

Implications

Deduplication halves the number of emails to process, which halves both processing time and cost.