Joshua Krasnogorov

Martin Cenek

CS-429

6 April 2025

Assignment 6 Write-Up

The “spam assassin” feature vector performs phenomenally well considering the limited number of vectors. With a batch of just 100 emails, it was able to correctly identify 87.9% of spam emails. That number went up to 96.3% with a training set of 12,400. Considering the virtually infinite dataset we have access to it’s no surprise that current spam filters are incredibly reliable. Since increasing the partition size by 100 wouldn’t lead to me approaching the 12400/3100 split, I used the np.linspace tool to create equal partitions from 100 to 12400.

train size	precision	recall	accuracy
100	0.9941	0.8790	0.9454
1394	0.9881	0.9521	0.9747
2689	0.9840	0.9607	0.9770
3984	0.9867	0.9623	0.9792
5278	0.9849	0.9612	0.9787
6573	0.9824	0.9824	0.9778
7868	0.9840	0.9840	0.9777
9163	0.9812	0.9559	0.9790
10457	0.9808	0.9636	0.9837
11752	0.9618	0.9641	0.9834
12400	0.9536	0.9629	0.9861

Only the first batch and every other batch after (as well as the last batch) are recorded for brevity