Spam filtering is a huge task

Spam filtering proprietary dataset

- https://arxiv.org/pdf/0902.2206.pdf
- 0.4 million users
- 3.2 million letters
- 40 million unique words

Let's say we map each token to index using hash function ϕ

- $\phi(x) = \text{hash}(x) \,\%\, 2^b$
- For $b = 22$ we have 4 million features
- That is a huge improvement over 40 million features
- It turns out it doesn't hurt the quality of the model