Statistics for Machine Learning and Big Data

An Introduction

Part 5: Spam

Jeff Abrahamson

7 April 2015

Bogofilter

- Paul Graham, *A Plan for Spam*
- Bag of words
- Bayes theorem

http://www.paulgraham.com/spam.html

Bogofilter

- Want: map from word to Pr email containing it is spam
- Bias to avoid false positives
- Minimum occurrence threshold
- Top 15 spam and ham words

Break

Questions?

`purple.com/talk-feedback`