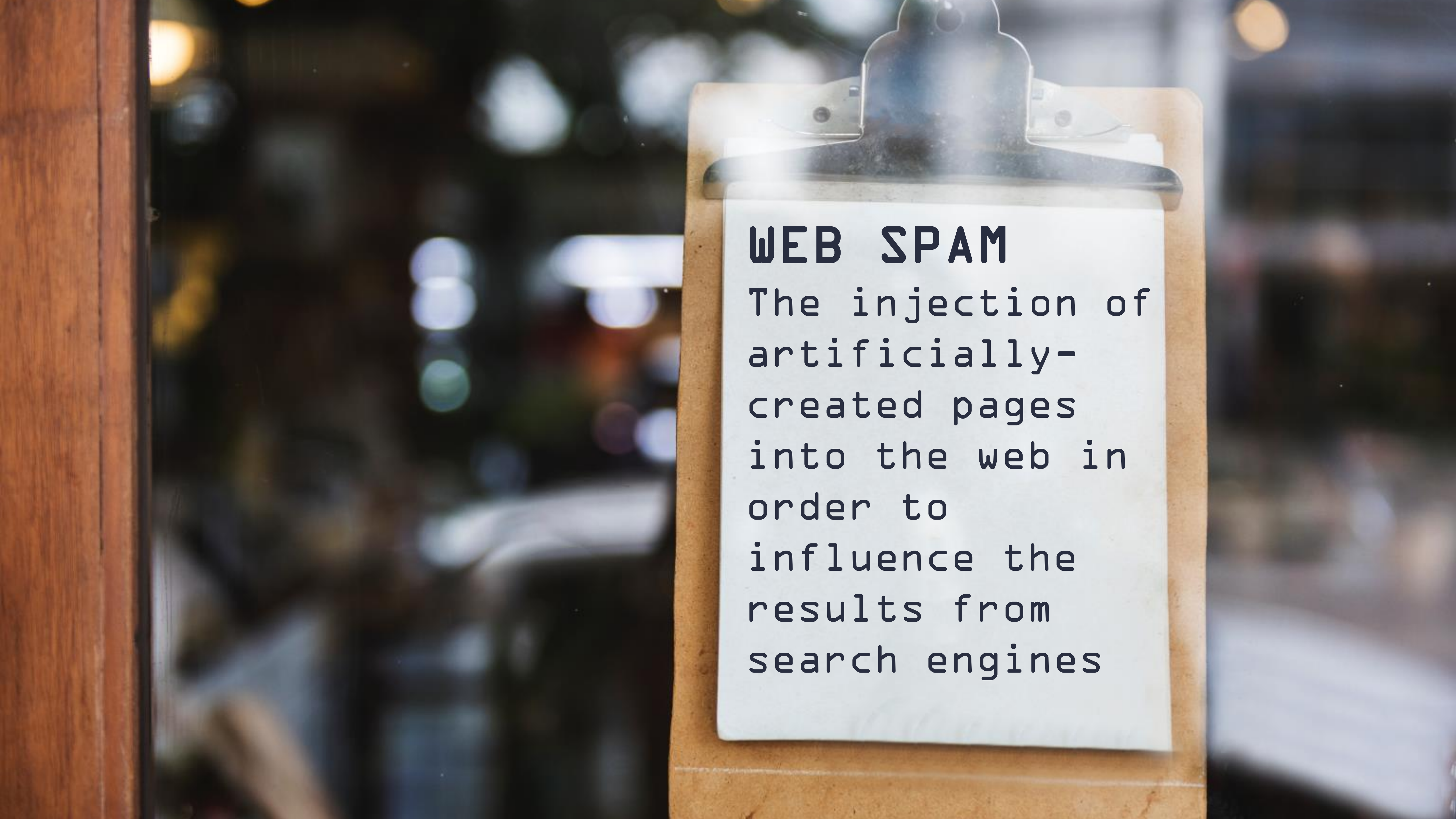
DETECTING WEB PAGES THROUGH CONTENT ANALYSIS

Viktoriya Bodnar

UniGe 2018



WEB SPAM

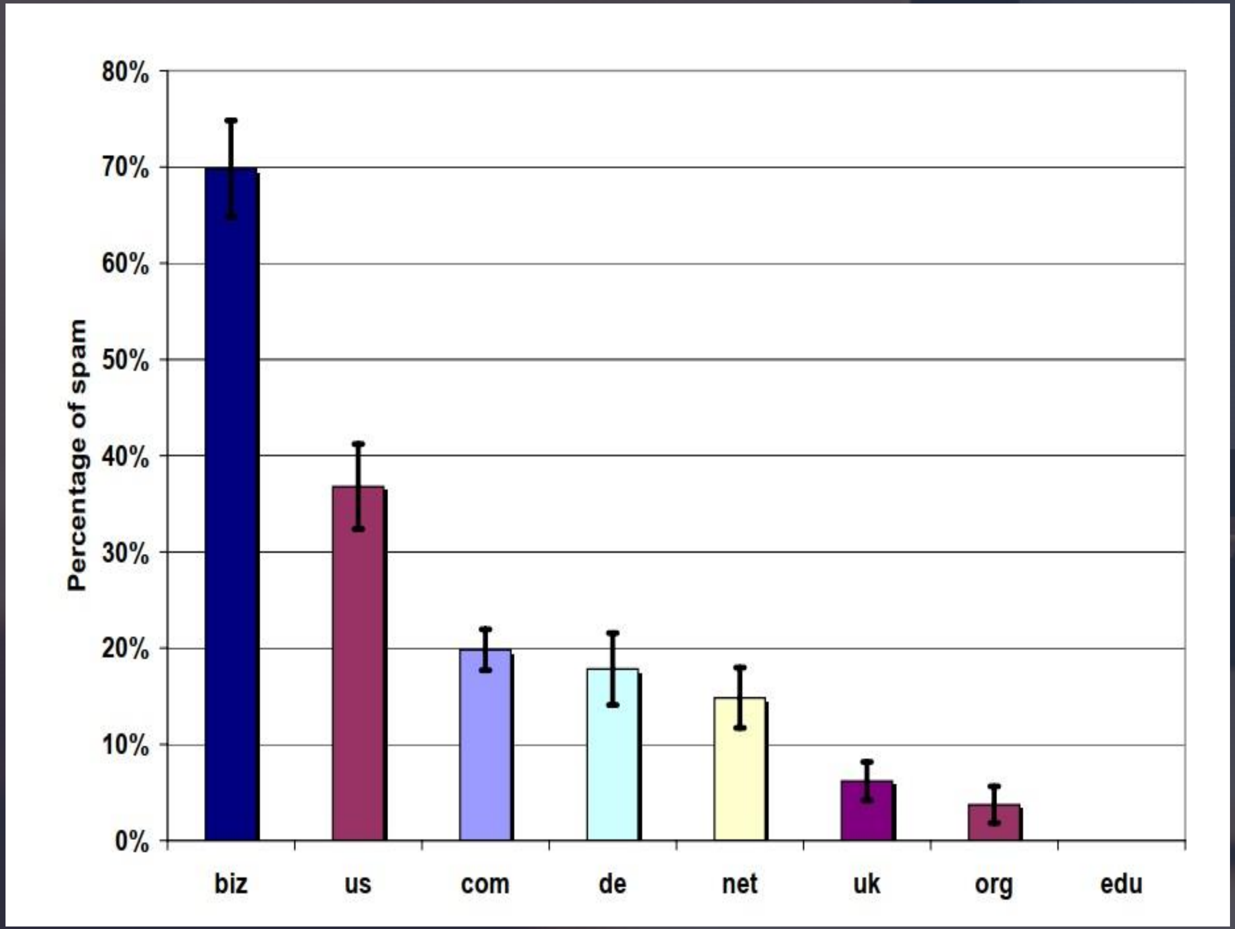
The injection of artificially-created pages into the web in order to influence the results from search engines

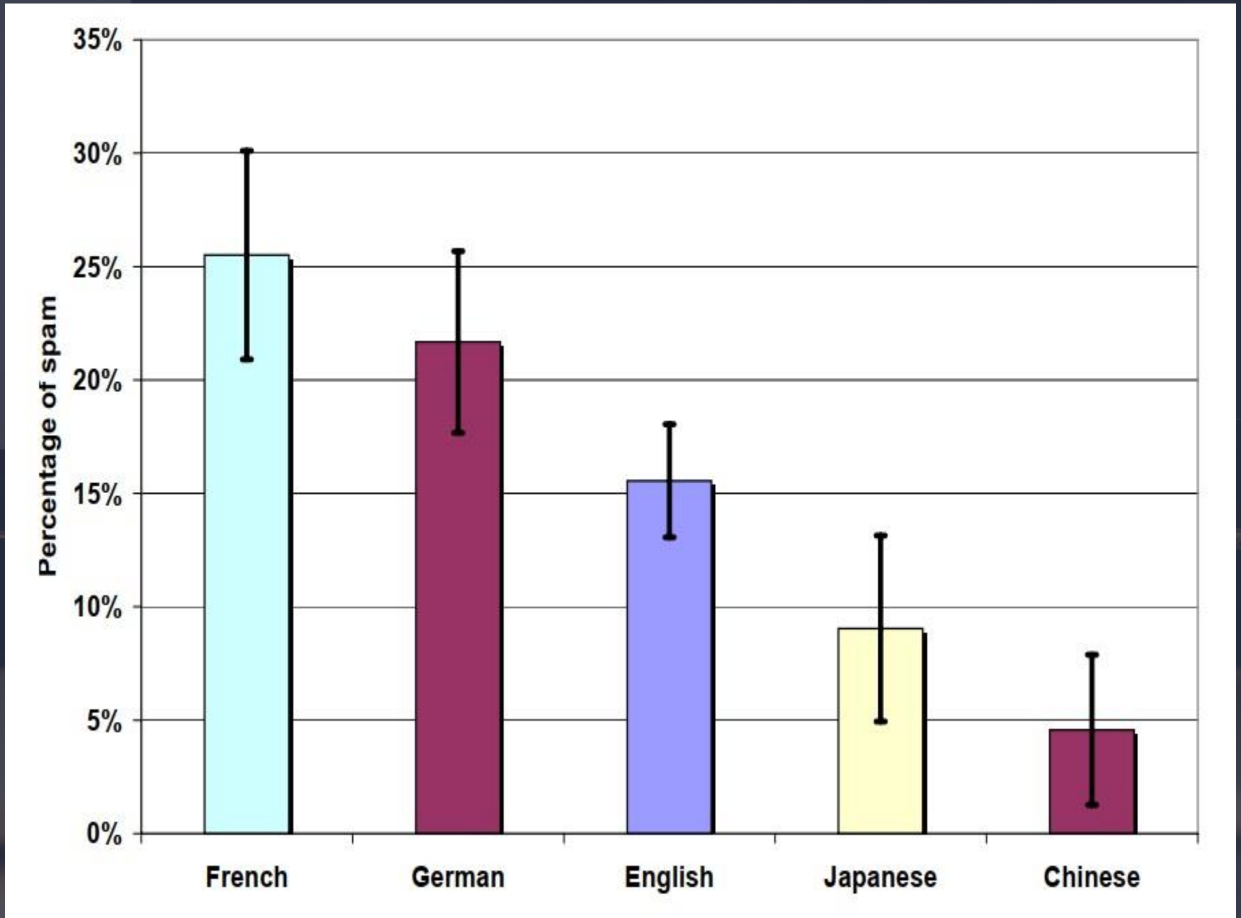
Alexandros Ntoulas, Marc Najork
Mark Manasse, Dennis Fetterly
DATASET

- 105, 484, 446 web pages
- MSN Search crawler
- August 2004
- uniform random sample 17, 168 English-written pages
- 2, 364 pages (13.8%) -> spam
- 14, 804 (86.2%) -> non-spam

Manually labeled

SOME STATISTICS



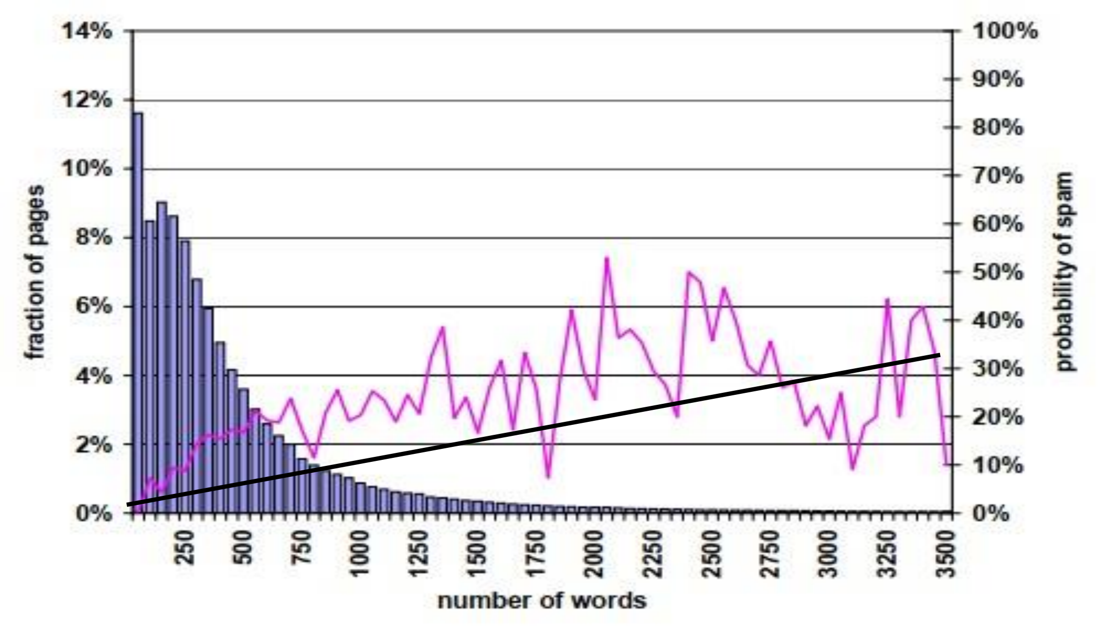


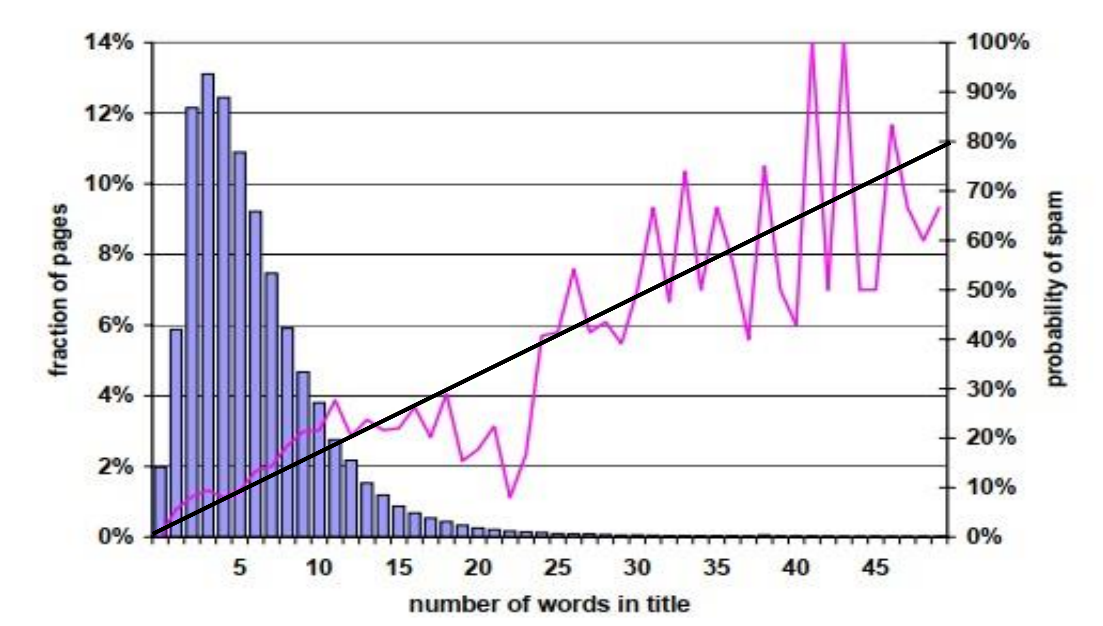
CONTENT-BASED SPAM DETECTION

1)number of words -> KEYWORDS STUFFING

- Number of words in the page

- Number of words in the page title

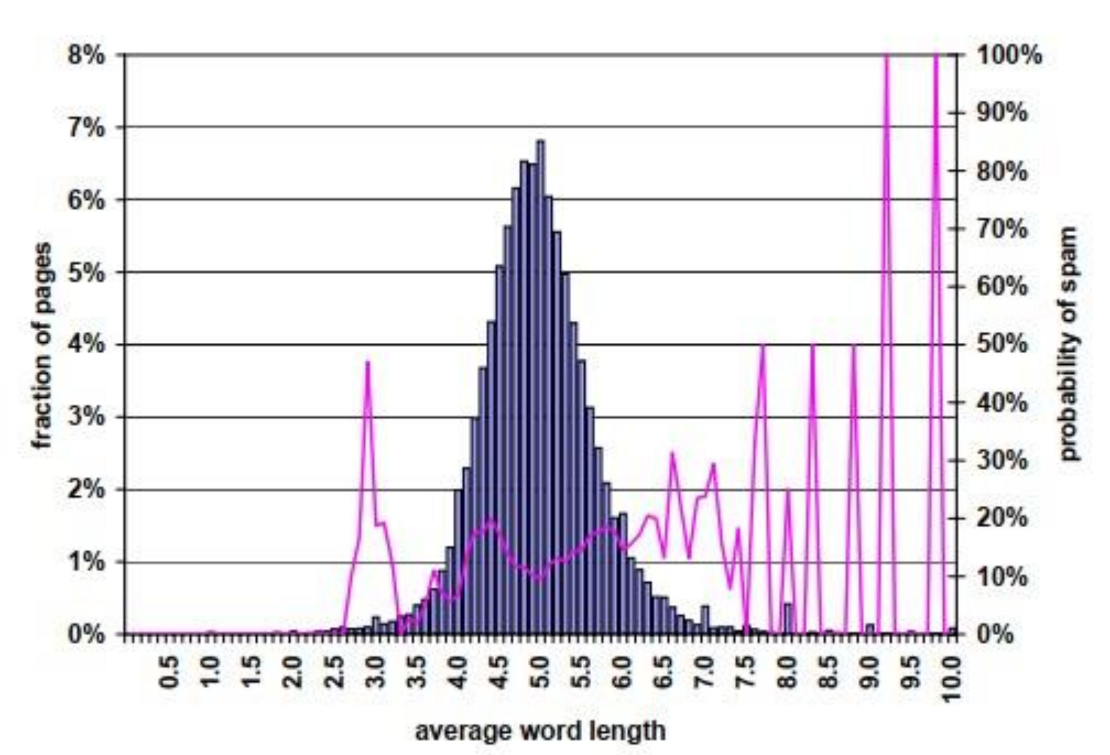




clear correlation between word count and prevalence of spam

CONTENT-BASED SPAM DETECTION
2)average words length

- composite words: freepictures, freedownload…
- target -> misspelled queries



avg word length 9 -> spam
language sensitive

CONTENT-BASED SPAM DETECTION

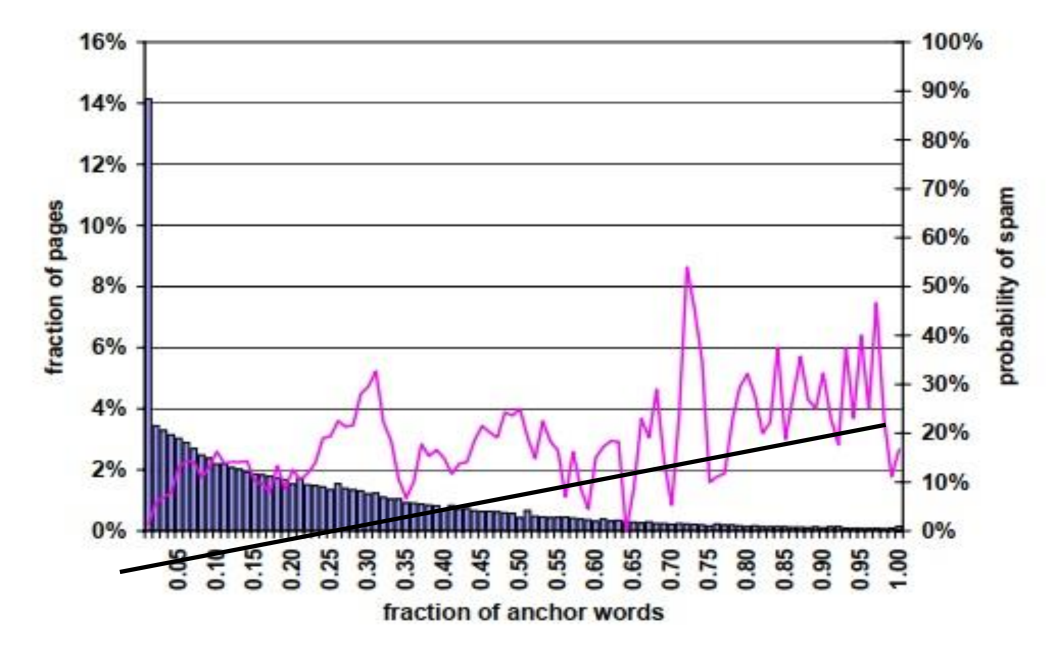
3)fraction of anchor/visible text

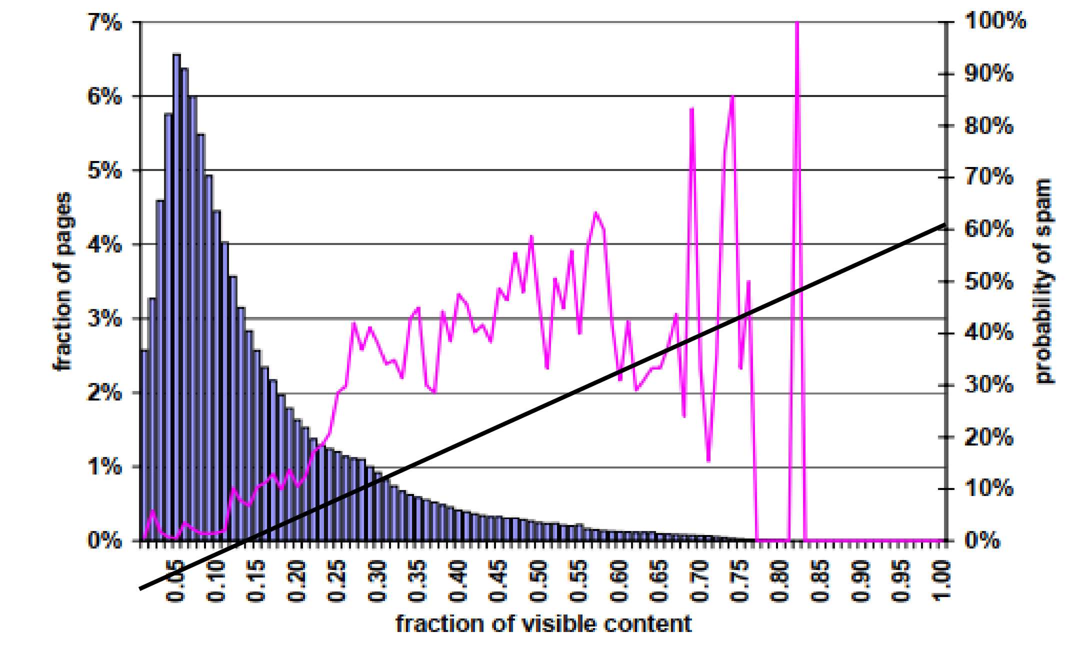
- <a href=“”> Anchor text

catalog of links

- comments, alt, meta

markup vs visible

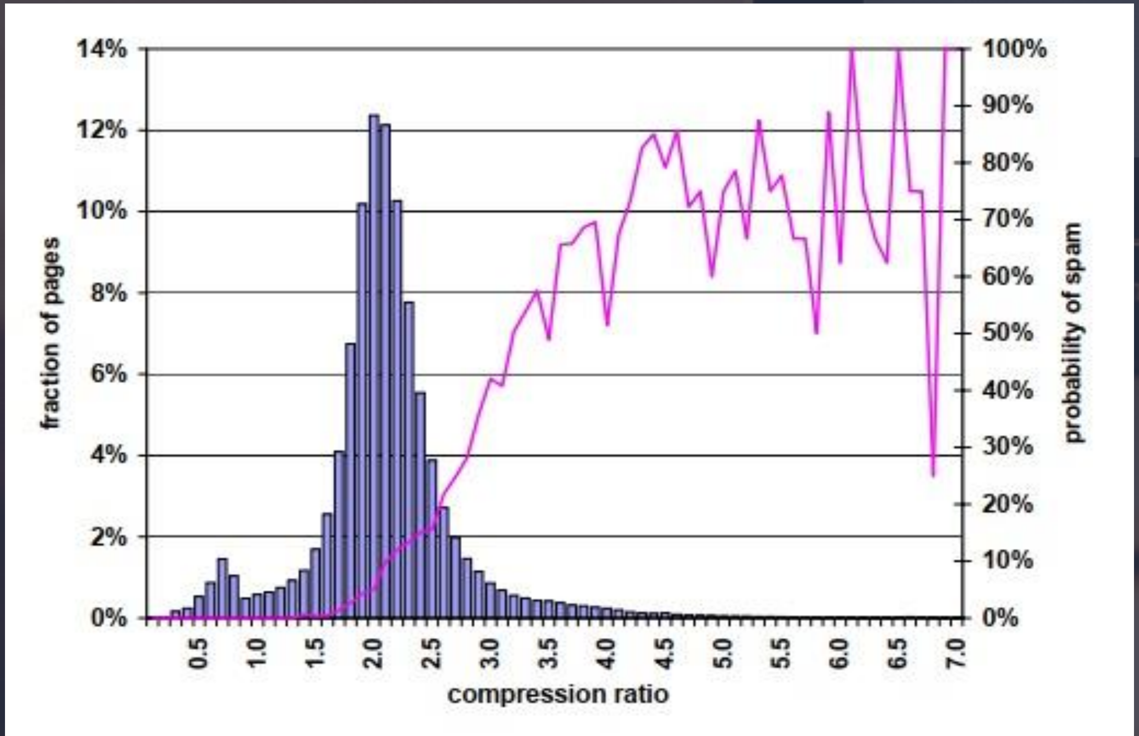




But! more code -> more probability of spam

CONTENT-BASED SPAM DETECTION
3)compressibility

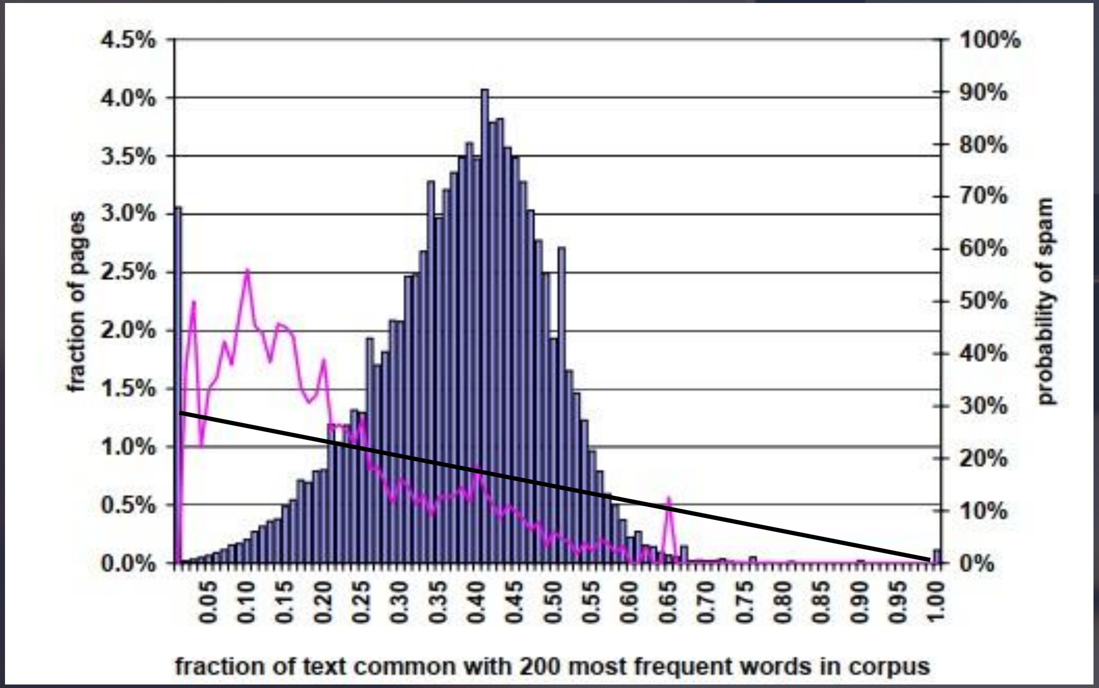
- redundant content
- compression ratio

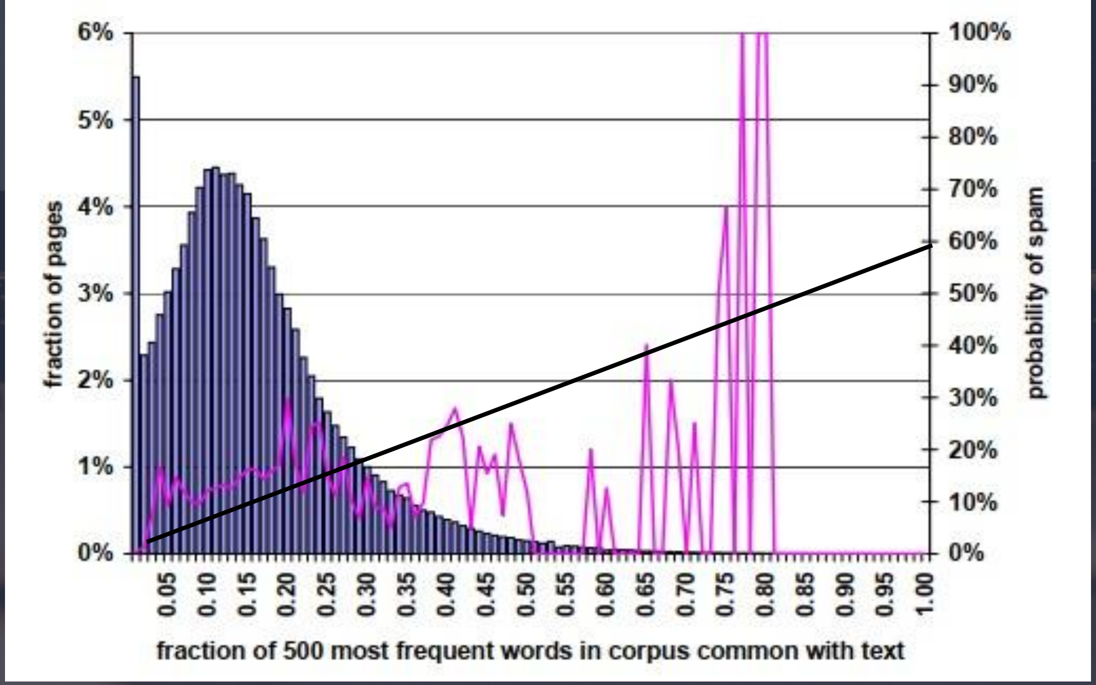


CONTENT-BASED SPAM DETECTION

3)fraction of popular words

- Fraction of most popular words within the page
- Fraction of globally popular words in the page





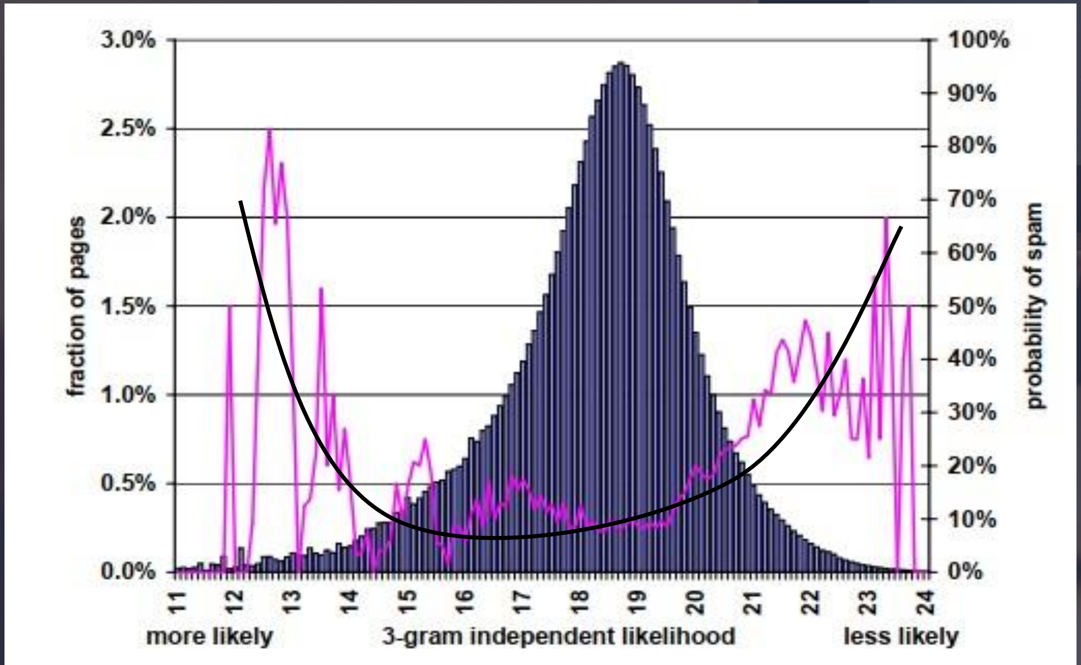
score 1 ⬅ "a" is one of the 500 most popular
a web page with just one word "a" ➡ score 1/500

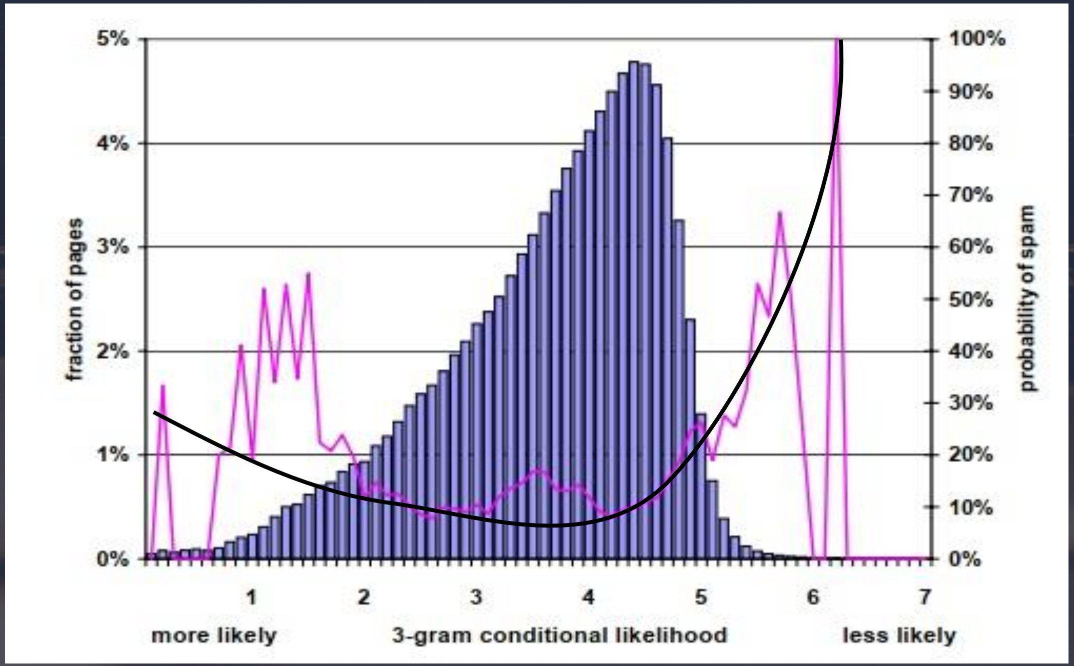
CONTENT-BASED SPAM DETECTION

4)n-gram likelihood

- independent
- conditional

grammatical and semantic correctness

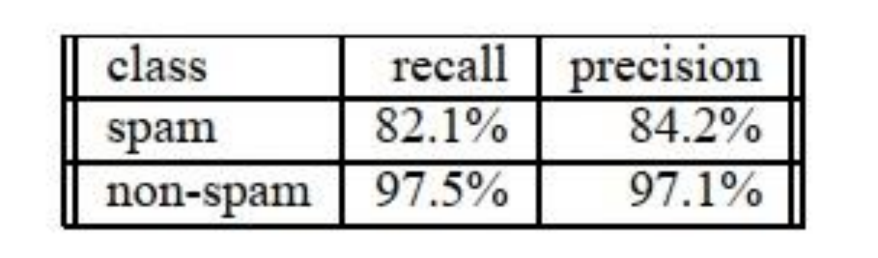




high value --> infrequently occurring n-grams

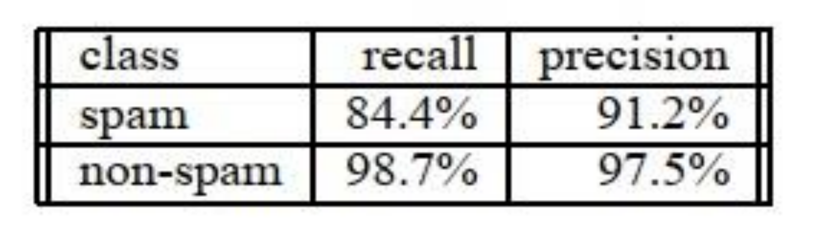
CLASSIFIER
to combine heuristics

- simple decision tree classifier
- ten fold cross validation
- 95.4% -> correctly
- 4.6% -> incorrectly



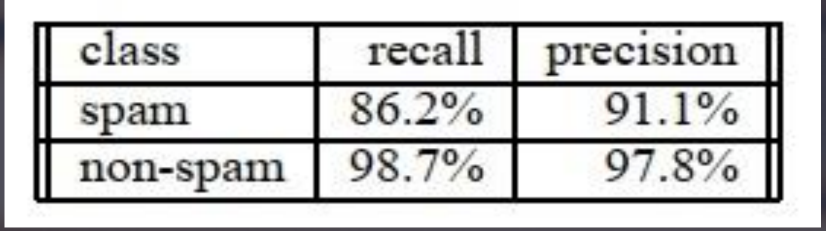
class	recall	precision
spam	82.1%	84.2%
non-spam	97.5%	97.1%

- composite classifier
- bagging -> majority of votes from N classifiers

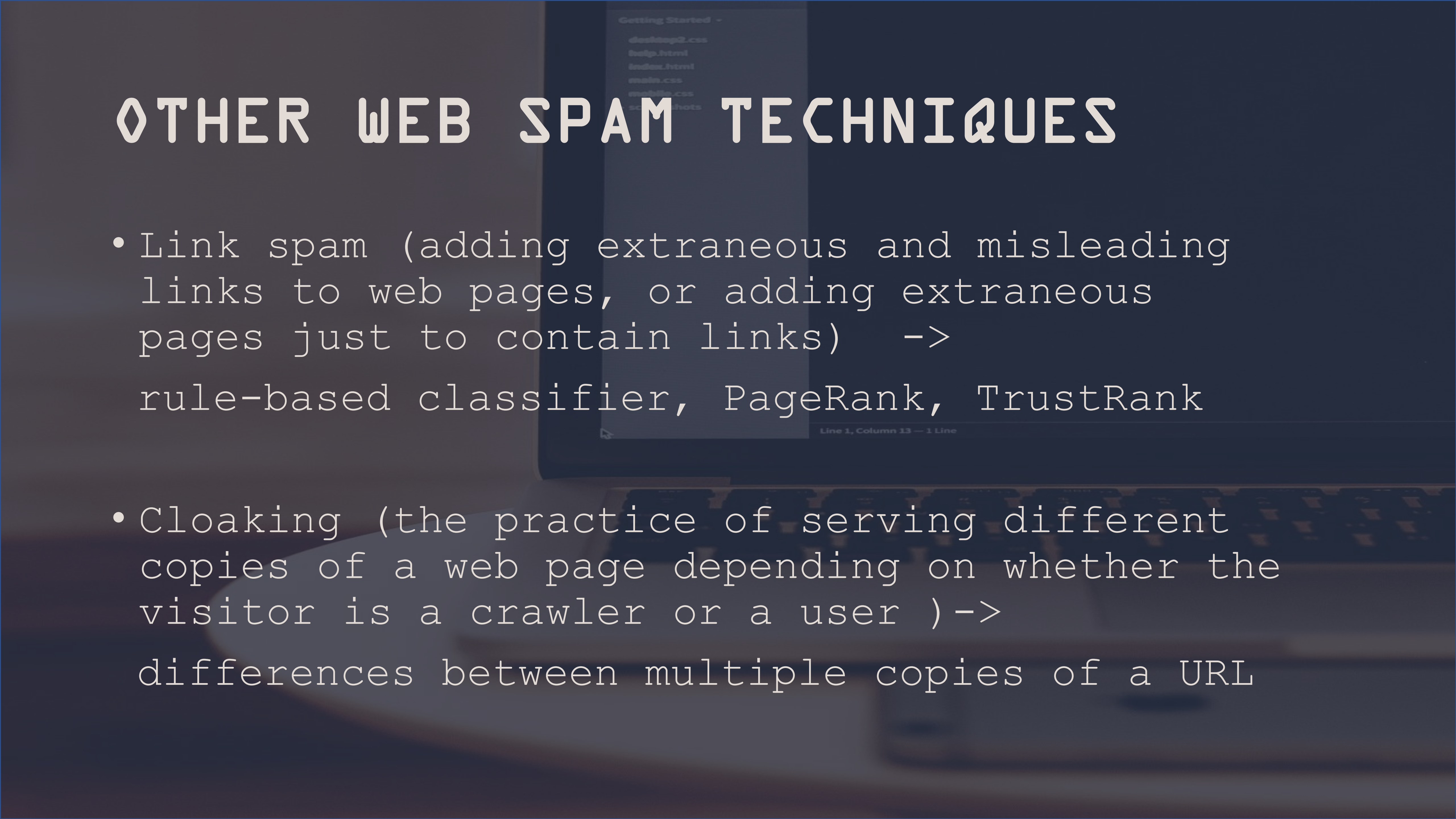


class	recall	precision
spam	84.4%	91.2%
non-spam	98.7%	97.5%

- boosting -> weighted sum of “votes” from N classifiers



class	recall	precision
spam	86.2%	91.1%
non-spam	98.7%	97.8%



OTHER WEB SPAM TECHNIQUES

- Link spam (adding extraneous and misleading links to web pages, or adding extraneous pages just to contain links) ->

  rule-based classifier, PageRank, TrustRank

- Cloaking (the practice of serving different copies of a web page depending on whether the visitor is a crawler or a user)->

  differences between multiple copies of a URL

BE AWARE OF SPAM PAGES

THANK YOU FOR ATTENTION!