Do Not Crawl In The DUST: Different URLS, Similar Text

Uri Schonfeld, Ziv Bar-Yossef, and Idit Keidar (Technion)

DUST Free URL List

http://site/story?id=2300$
http://site/story?id=1400$
http://site/money/story?id=340$
http://site/story?id=340$

Dusty URL List

http://site/story_2300$
http://site/story?id=2300$
http://site/story?id=2300&view=print$
http://site/story?id=1400$
http://site/story?id=1400&view=print$
http://money.site/story?id=340$
http://site/money/story?id=340$
http://site/story?id=340$

DustBuster Flow

Likely DUST Rules

story_ ↔ story?id=
&view=print$ ↔ $
money.site ↔ site/money
site/ ↔ money. site/

Detection

Training URL List

http://site/story_130$
http://site/story?id=130$
http://site/story_571$
http://site/story?id=571$
http://site/story?id=130&view=print$
http://money.site/story?id=571$
http://site/money/story?id=571$
http://site/story?id=571$

Sampling

Validation

Valid DUST Rules

story_ ➡ story?id=
&view=print$ ➡ $
money. site ➡ site/money

Dusty URL List

http://site/story_2300$
http://site/story?id=2300$
http://site/story?id=2300&view=print$
http://site/story?id=1400$
http://site/story?id=1400&view=print$
http://money.site/story?id=340$
http://site/money/story?id=340$
http://site/story?id=340$

Canonization1

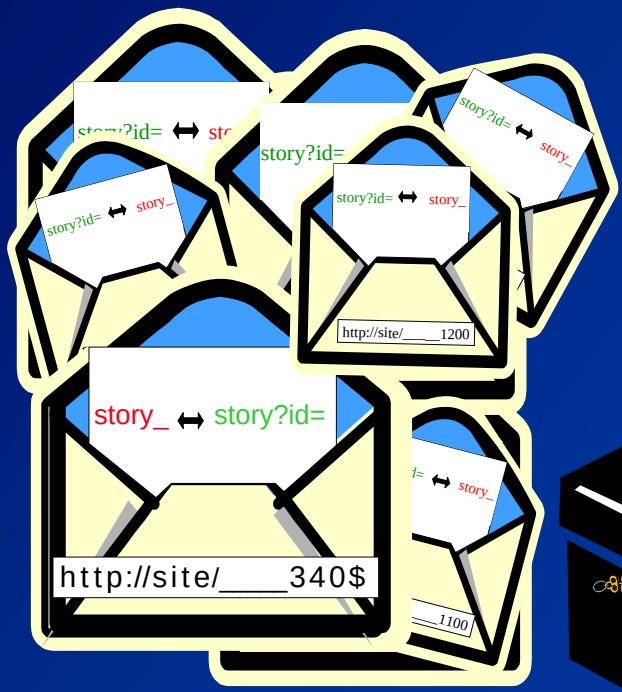
DUST Free URL List

http://site/story?id=2300$
http://site/story?id= 1400$
http://site/money/story?id=340$
http://site/story?id=340$

Why Does DustBuster Work?

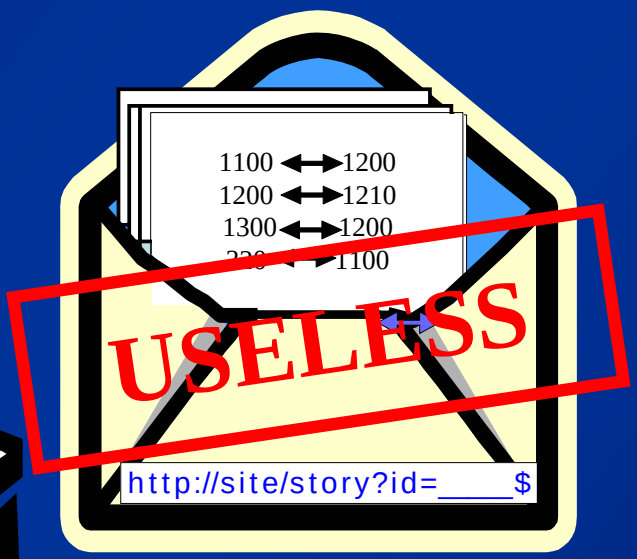
Large Support Principle

Valid Rules Get Many Votes



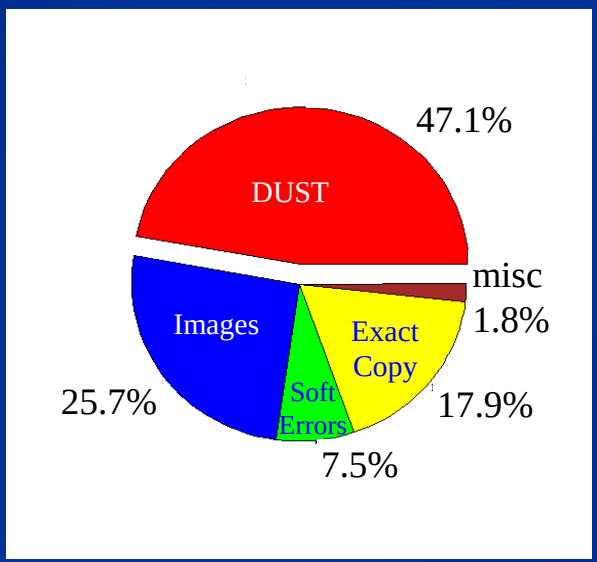
Fat Envelopes Principle

Fat Envelopes Are Useless

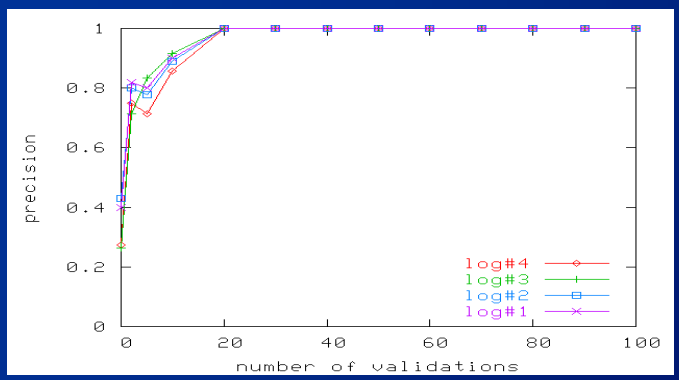


Results From Real Web Sites

DUST Distribution



Precision vs. Validations



Precision@k with and without size