Thursday
Ronald M. Kaplan
Agenda

• HW5 concepts
  · Merge
    ▪ Cosine similarity
• Activity: improve a bad IR system
• Activity: search engine optimization
HW5 (due Friday)

• A search engine
  • Inverted index
  • Boolean retrieval
  • TF-IDF weighting
  • Rank with cosine similarity
Agenda

• HW5 concepts
  ▪ Merge
  ▪ Cosine similarity
• Activity: improve a bad IR system
• Activity: search engine optimization
Boolean retrieval

• In video: merging for “cat AND dog”
• Warmup: merging for “cat OR dog”
• Runtime?
• Question: Runtime for “cat AND NOT dog”? 
Agenda

• HW5 concepts
  - Merge
  - Cosine similarity
• Activity: improve a bad IR system
• Activity: search engine optimization
Cosine similarity

- Documents as vectors
- Document similarity as vector distance
- Not Euclidean distance
- Euclidean distance ignoring lengths = cosine similarity
- Why is the formula $q \cdot d / ||q|| \ ||d||$?
Agenda

• HW5 concepts
  - Merge
  - Cosine similarity
• Activity: improve a bad IR system
• Activity: search engine optimization
What's wrong here?

- **Query**: “the cat in the hat”

- **Top result**: ...at the end of the day, most stakeholders in the room agreed that the project in Dover is what broke the bank...

- **Result 2**: ...one of the songs in the album “in situ” by the band “the fray”...
What's wrong here?

• **Query**: “the cat in the hat”

• **Top result**: ...at the end of the day, most stakeholders in the room agreed that the project in Dover is what broke the bank...

• **Result 2**: ...one of the songs in the album “in situ” by the band “the help”...

Agenda

• HW5 concepts
  - Merge
  - Cosine similarity
• Activity: improve a bad IR system
• Activity: search engine optimization
SEO contest

• Changing your document to improve its rank
• Warm up question: you know my query.
SEO contest

• Secret query

• But you know that I want to **pick a wedding present**

• **Task**: Write a document designed to outrank others

• Spammy documents are welcome

• Upload to
  
  /afs/ir/class/cs124/sections/section5/seo-docs/Your_Name(s).txt
  
  (can use [http://afs.stanford.edu](http://afs.stanford.edu))
Thank you