Information Retrieval

Natural Language Processing: Jordan Boyd-Graber
University of Maryland

OVERVIEW

Slides adapted from Jimmy Lin
IR is worth a lot of money . . .
Search Engines Rule the Map
Why we’re starting here

- Simple to understand (hard to implement)
- Important to the world
- Building block for course project
Prerequisites

- Search a “collection” of documents
- Each document contains terms (words)
- Users create queries
General Problem

User

“tragic romance”

Author

“star-crossed lovers”
What’s a word?

We’ll talk more about this later ...
What’s a word?

We'll talk more about this later . . . assume we know the answer.
Bag of Words

**McDonald's slims down spuds**

Fast-food chain to reduce certain types of fat in its french fries with new cooking oil.

NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

But does that mean the popular shoestring fries won't taste the same? The company says no. "It's a win-win for our customers because they are getting the same great french-fry taste along with an even healthier nutrition profile," said Mike Roberts, president of McDonald's USA.

But others are not so sure. McDonald's will not specifically discuss the kind of oil it plans to use, but at least one nutrition expert says playing with the formula could mean a different taste.

Shares of Oak Brook, Ill.-based McDonald's (MCD: down $0.54 to $23.22, Research, Estimates) were lower Tuesday afternoon. It was unclear Tuesday whether competitors Burger King and Wendy's International (WEN: down $0.80 to $34.91, Research, Estimates) would follow suit. Neither company could immediately be reached for comment.

...
Boolean Retrieval

- Users express queries as a Boolean expression
  - AND, OR, NOT
  - Can be arbitrarily nested
  - Retrieval is based on the notion of sets
- Any given query divides the collection into two sets: retrieved, not-retrieved
  - Pure Boolean systems do not define an ordering of the results
Strengths and Weaknesses

- **Strengths**
  - Precise, if you know the right strategies
  - Precise, if you have an idea of what you’re looking for
  - Implementations are fast and efficient

- **Weaknesses**
  - Users must learn Boolean logic
  - Boolean logic insufficient to capture the richness of language
  - No control over size of result set: either too many hits or none
  - When do you stop reading? All documents in the result set are considered “equally good”
  - What about partial matches? Documents that “don’t quite match” the query may be useful also