



Information Retrieval

Natural Language Processing: Jordan
Boyd-Graber
University of Maryland

OVERVIEW

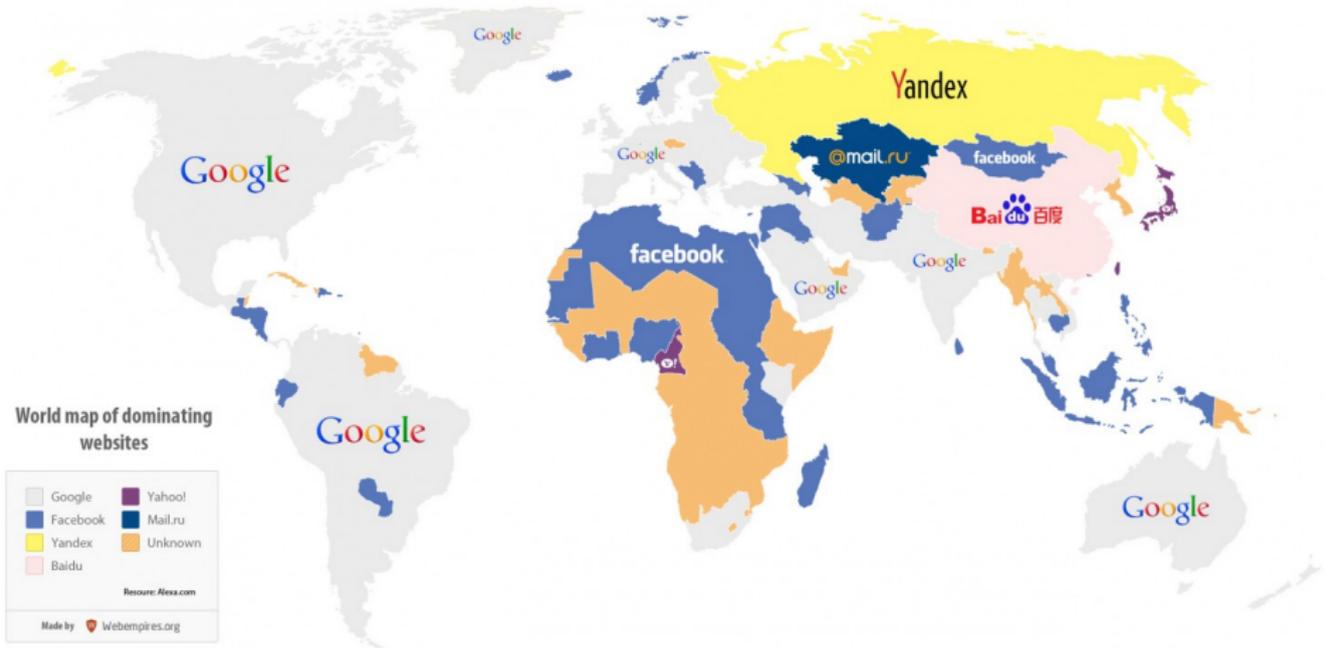
Slides adapted from Jimmy Lin

Google™

Google Search

I'm Feeling Lucky

IR is worth a lot of money . . .



Search Engines Rule the Map

Why we're starting here

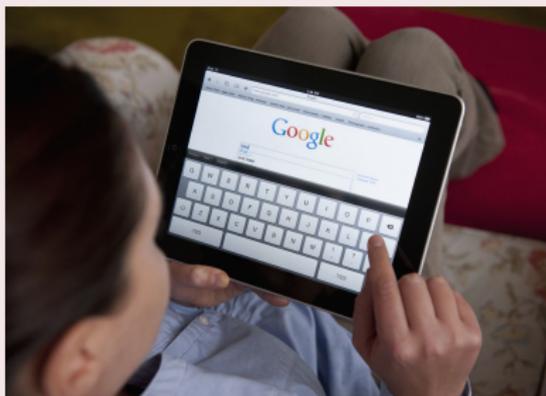
- Simple to understand (hard to implement)
- Important to the world
- Building block for course project

Prerequisites

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

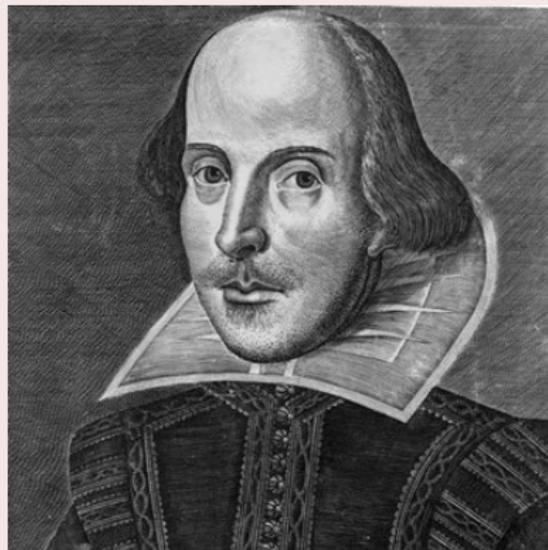
General Problem

User



“tragic romance”

Author



“star-crossed lovers”

What's a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

وقال مارك ريجيف - الناطق باسم
الخارجية الإسرائيلية - إن شارون قبل
الدعوة وسيقوم للمرة الأولى بزيارة
تونس، التي كانت لفترة طويلة المقر
الرسمي لمنظمة التحرير الفلسطينية بعد خروجه من لبنان عام 2002 .

Выступая в Мещанском суде Москвы экс-глава ЮКОСа
заявил не совершал ничего противозаконного, в чем
обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ्रीसदी विकास
दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処...アーミテージ前副長官提言

조재영 기자= 서울시는 25일 이명박 시장이 "행정중심복합도시" 건설안
에 대해 "군대라도 동원해 막고싶은 심정"이라고 말했다는 일부 언론의
보도를 부인했다.

We'll talk more about this later . . .

What's a word?

天主教教宗若望保祿二世因感冒再度住進醫院。
這是他今年第二度因同樣的病因住院。

وقال مارك ريجيف - المناطق باسم
الخارجية الإسرائيلية - إن شارون قبل
الدعوة وسيقوم للمرة الأولى بزيارة
تونس، التي كانت لفترة طويلة المقر
الرسمي لمنظمة التحرير الفلسطينية بعد خروجها من لبنان عام 2005 .

Выступая в Мещанском суде Москвы экс-глава ЮКОСа
заявил не совершал ничего противозаконного, в чем
обвиняет его генпрокуратура России.

भारत सरकार ने आर्थिक सर्वेक्षण में वित्तीय वर्ष 2005-06 में सात फ्रीसदी विकास
दर हासिल करने का आकलन किया है और कर सुधार पर ज़ोर दिया है

日米連合で台頭中国に対処...アーミテージ前副長官提言

조재영 기자= 서울시는 25일 이명박 시장이 "행정중심복합도시" 건설안
에 대해 "군대라도 동원해 막고싶은 심정"이라고 말했다는 일부 언론의
보도를 부인했다.

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.

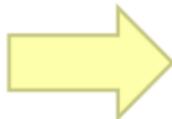
NEW YORK (CNN/Money) - McDonald's Corp. is cutting the amount of "bad" fat in its french fries nearly in half, the fast-food chain said Tuesday as it moves to make all its fried menu items healthier.

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.

...



"Bag of Words"

14 × McDonalds

12 × fat

11 × fries

8 × new

7 × french

6 × company, said, nutrition

5 × food, oil, percent, reduce,
taste, Tuesday

...

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