David **Milne** | Ian H. **Witten**

Knowledge-based Information Retrieval

with

Wikipedia

- Koru
- Wikipedia Link-based Measure
- Wikification

The University of Waikato | **New Zealand**

Limitations of search engines

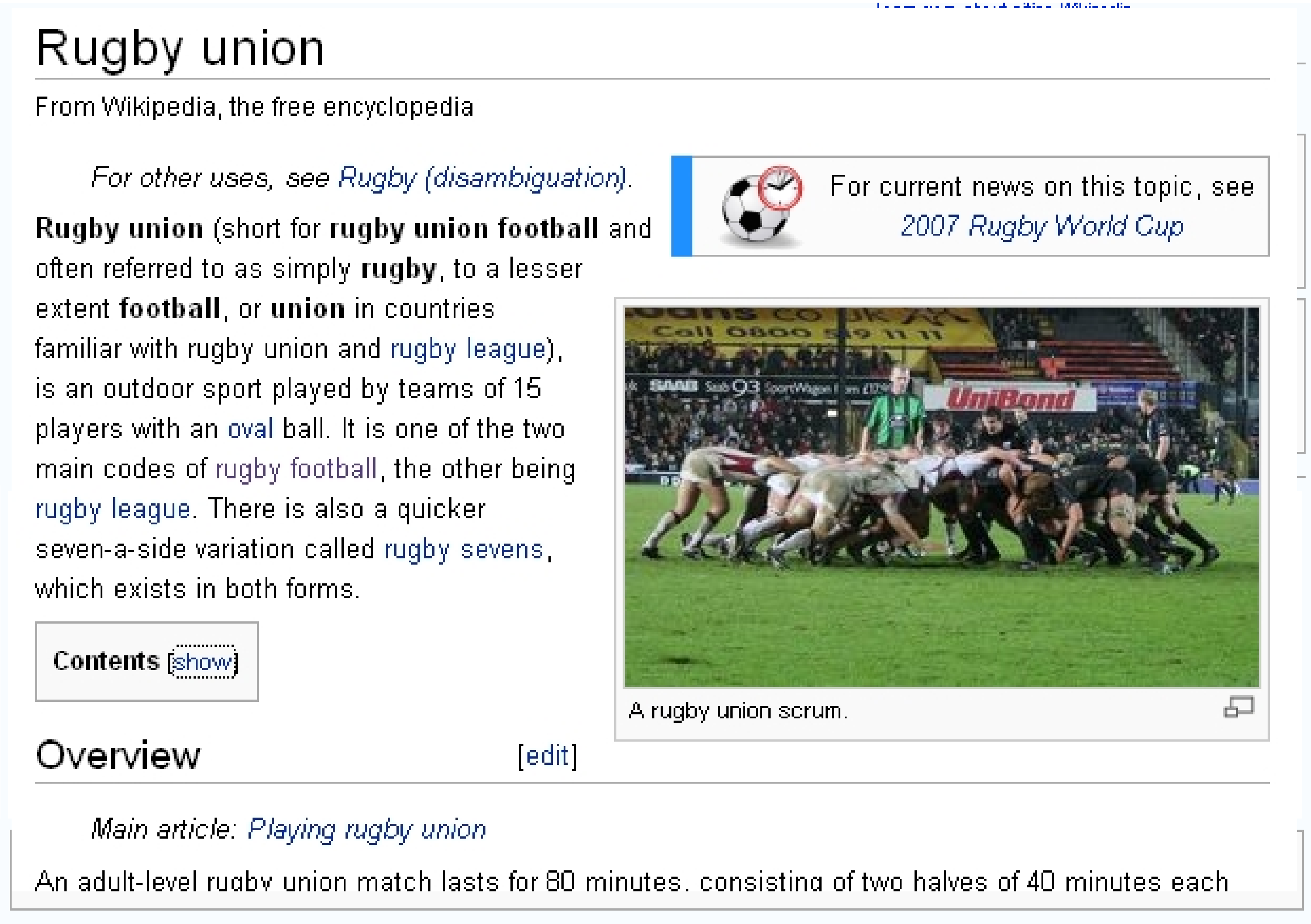
- “Search is not solved”
- Current search engines
  - don’t understand documents
  - don’t understand queries

Knowledge-based information retrieval

- Consult an external knowledge base
  - to find out what these characters mean
  - and proactively do stuff with them
- A fairly obvious, compelling idea
  - But one that hasn’t worked out
- We haven’t had the right knowledge base
  - Computers aren’t accurate enough
  - Humans aren’t quick enough

Wikipedia | as a knowledge base

- V
- H
- H



Rugby union

From Wikipedia, the free encyclopedia

For other uses, see Rugby (disambiguation).

For current news on this topic, see *2007 Rugby World Cup*

Rugby union (short for **rugby union football** and often referred to as simply **rugby**, to a lesser extent **football**, or **union** in countries familiar with rugby union and rugby league), is an outdoor sport played by teams of 15 players with an oval ball. It is one of the two main codes of rugby football, the other being rugby league. There is also a quicker seven-a-side variation called rugby sevens, which exists in both forms.

Contents [show]

A rugby union scrum.

Overview [edit]

Main article: Playing rugby union

An adult-level rugby union match lasts for 80 minutes, consisting of two halves of 40 minutes each

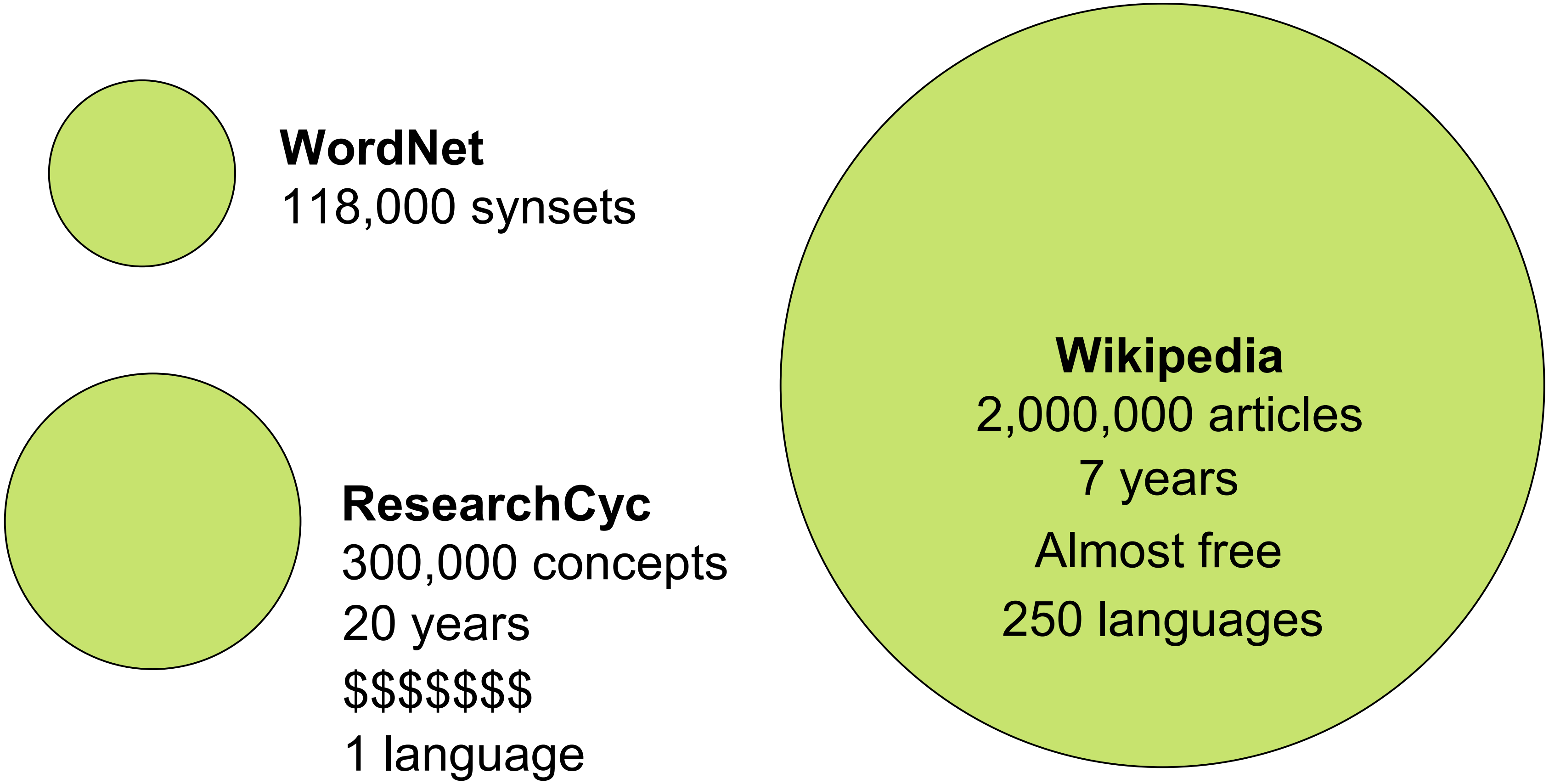
Wall

l sports

n rugby

New Zealand national rugby team

Wikipedia | as a knowledge base

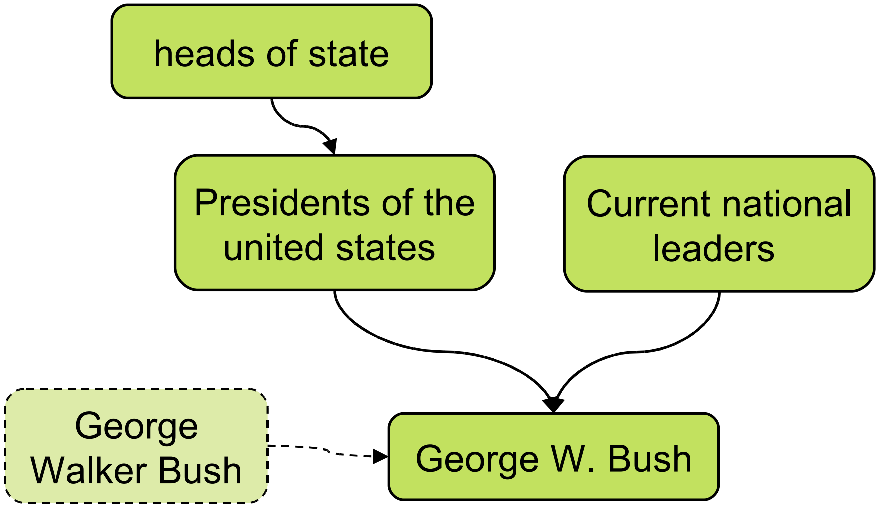


WordNet
118,000 synsets

ResearchCyc
300,000 concepts
20 years
$$$$$$$
1 language

Wikipedia
2,000,000 articles
7 years
Almost free
250 languages

Wikipedia | as a knowledge base



heads of state

Presidents of the united states

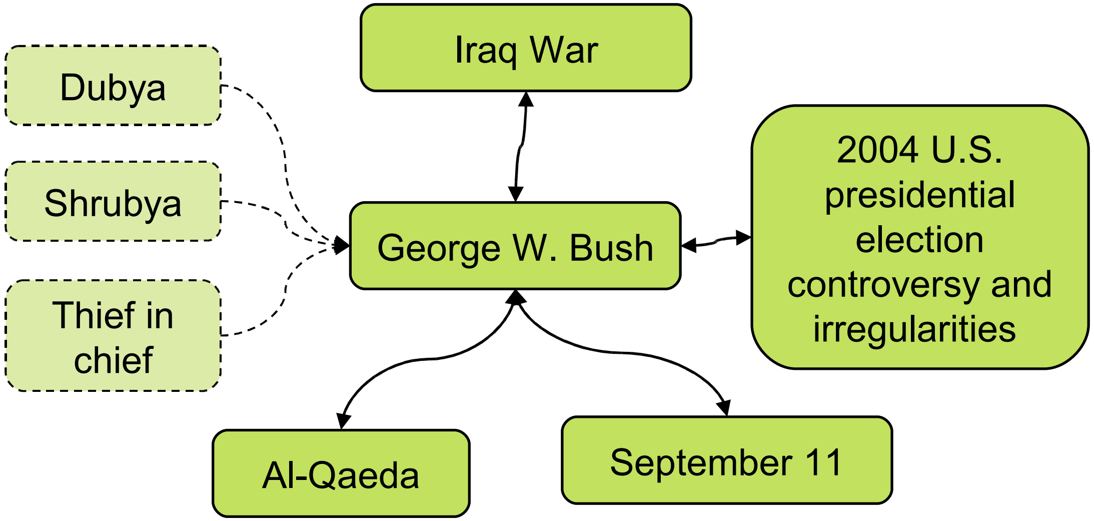
Current national leaders

George Walker Bush

George W. Bush

◄ **Formal structure**

Wikipedia ►



Dubya

Shrubya

Thief in chief

Iraq War

George W. Bush

2004 U.S. presidential election controversy and irregularities

Al-Qaeda

September 11

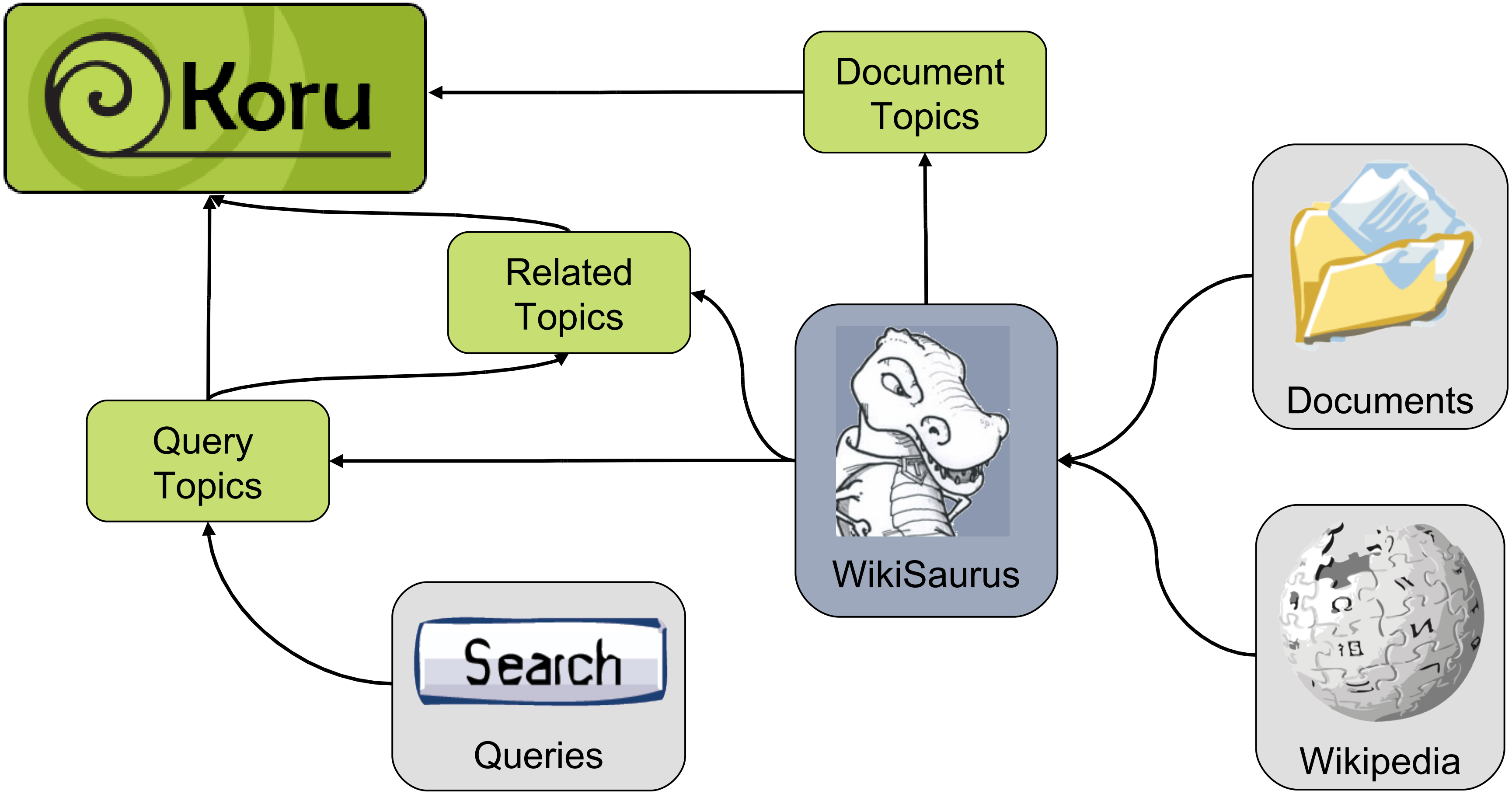
My research goals

“Wikipedia will provide significantly improved retrieval, as it is”

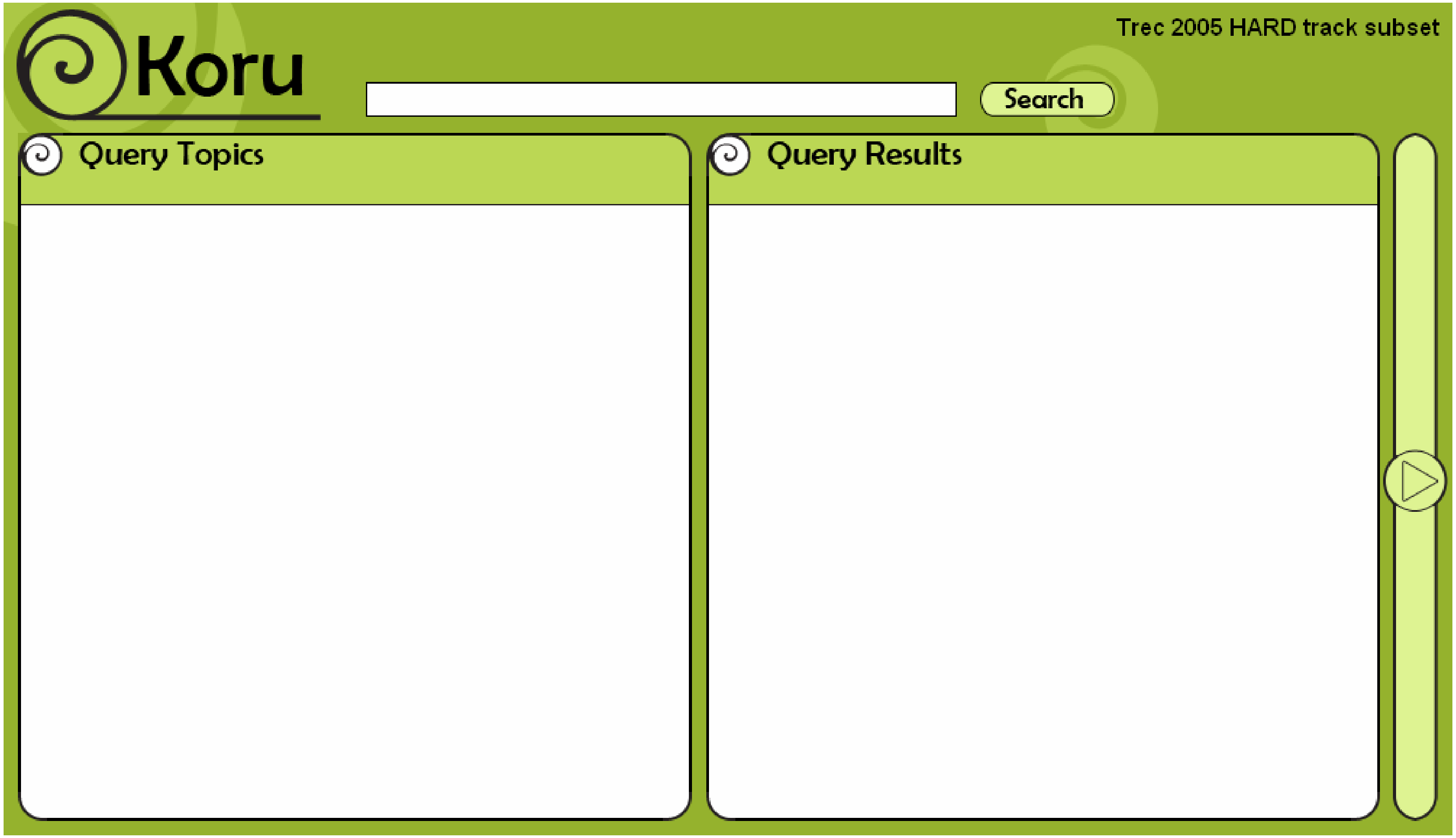
- We don’t need to make it “tidy”
- It’s not a question of sophisticated NLP or AI
- It’s more about HCI

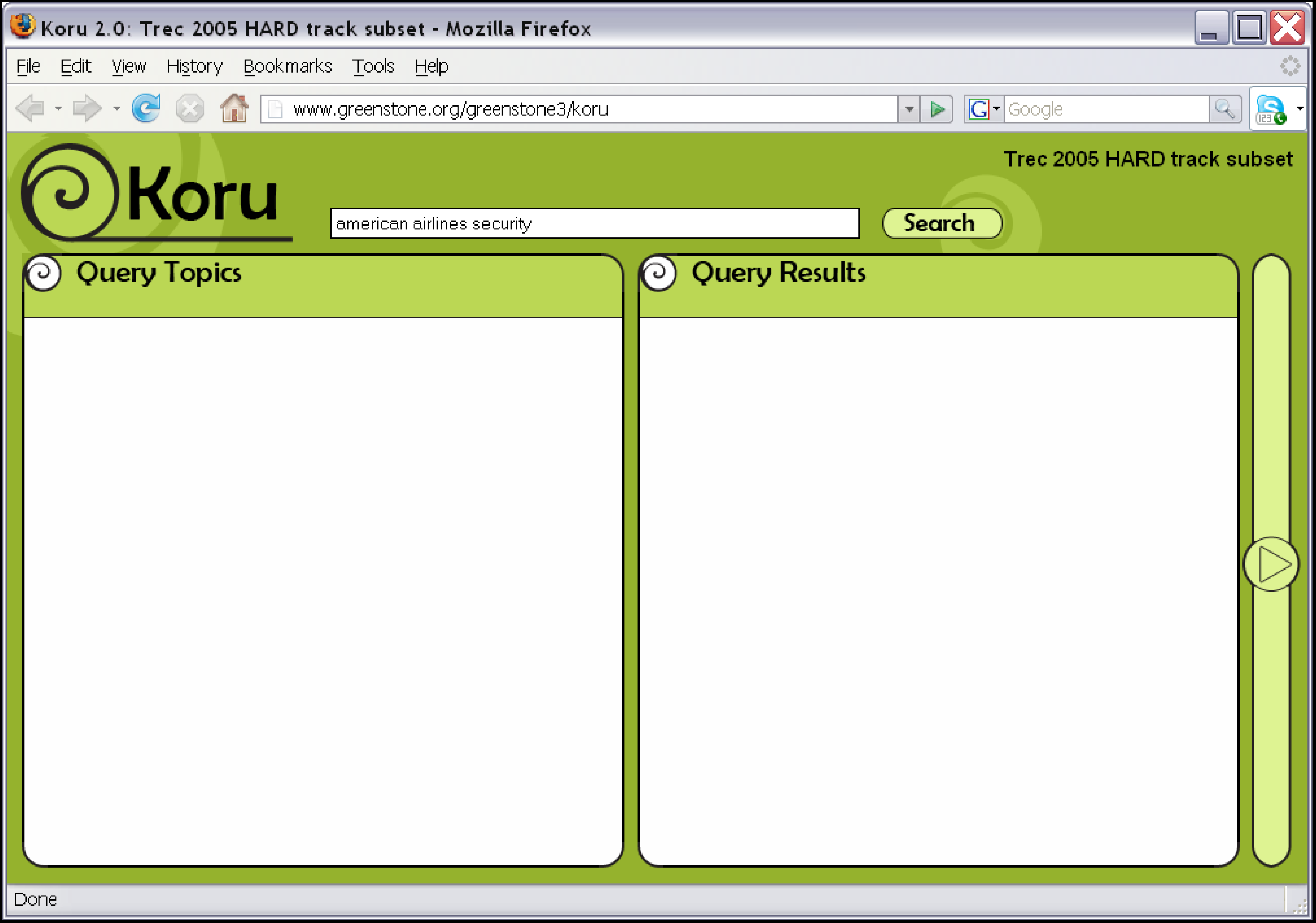
So lets make a search engine that consults Wikipedia, and find out!

Koru

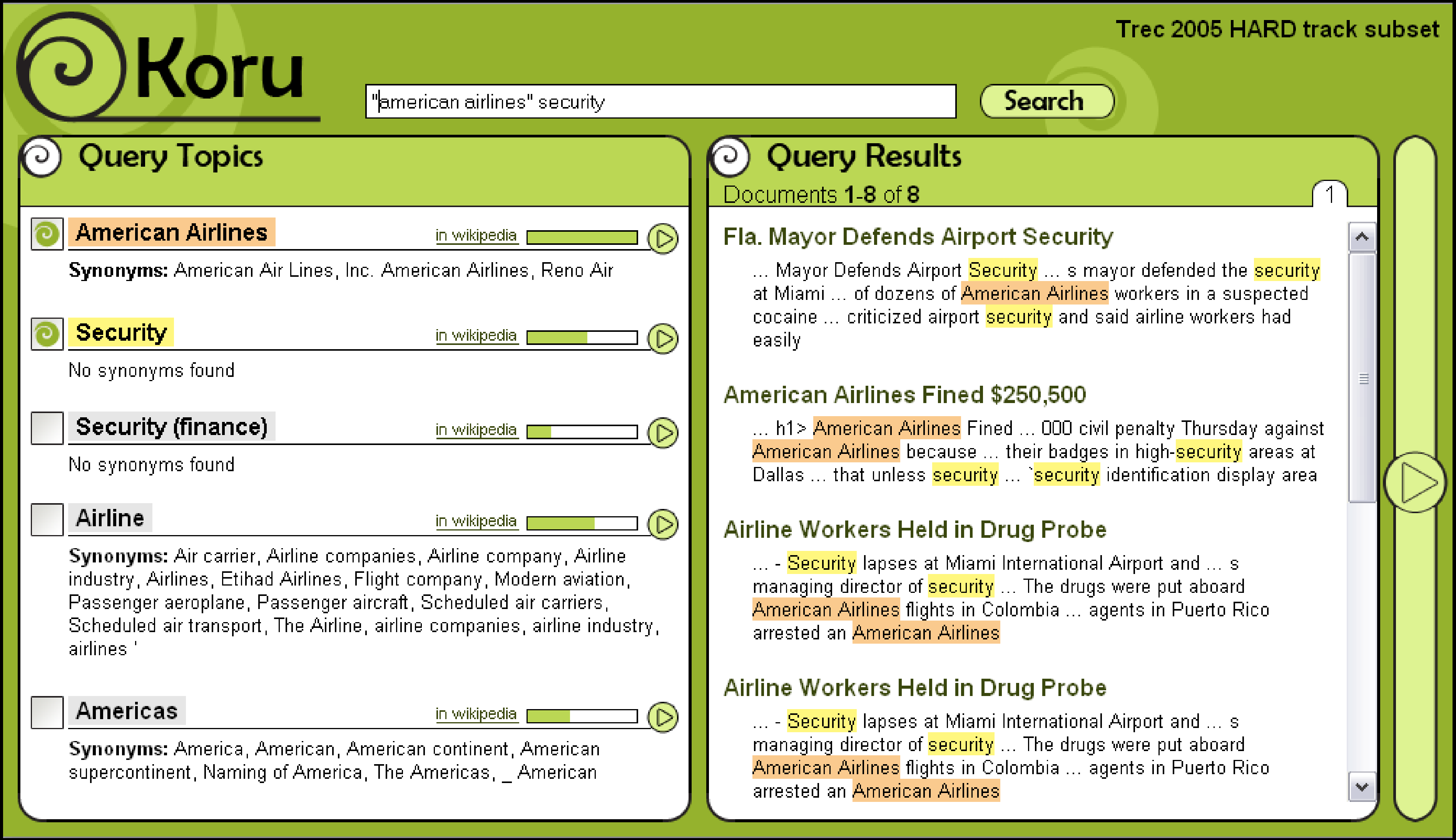

Koru
Document Topics
Related Topics
Query Topics
WikiSaurus
Documents
Wikipedia
Search
Queries

Koru | interface





Koru 2.0: Trec 2005 HARD track subset - Mozilla Firefox
File
Edit
View
History
Bookmarks
Tools
Help
www.greenstone.org/greenstone3/koru
Google
Koru
Trec 2005 HARD track subset
american airlines security
Search
Query Topics
Query Results
Done



"american airlines" security

Query Topics

☑ **American Airlines** — in wikipedia

Synonyms: American Air Lines, Inc. American Airlines, Reno Air

☑ **Security** — in wikipedia

No synonyms found

□ **Security (finance)** — in wikipedia

No synonyms found

□ **Airline** — in wikipedia

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

□ **Americas** — in wikipedia

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Query Results

Documents **1-8** of **8**

Fla. Mayor Defends Airport Security

... Mayor Defends Airport Security ... s mayor defended the security at Miami ... of dozens of American Airlines workers in a suspected cocaine ... criticized airport security and said airline workers had easily

American Airlines Fined $250,500

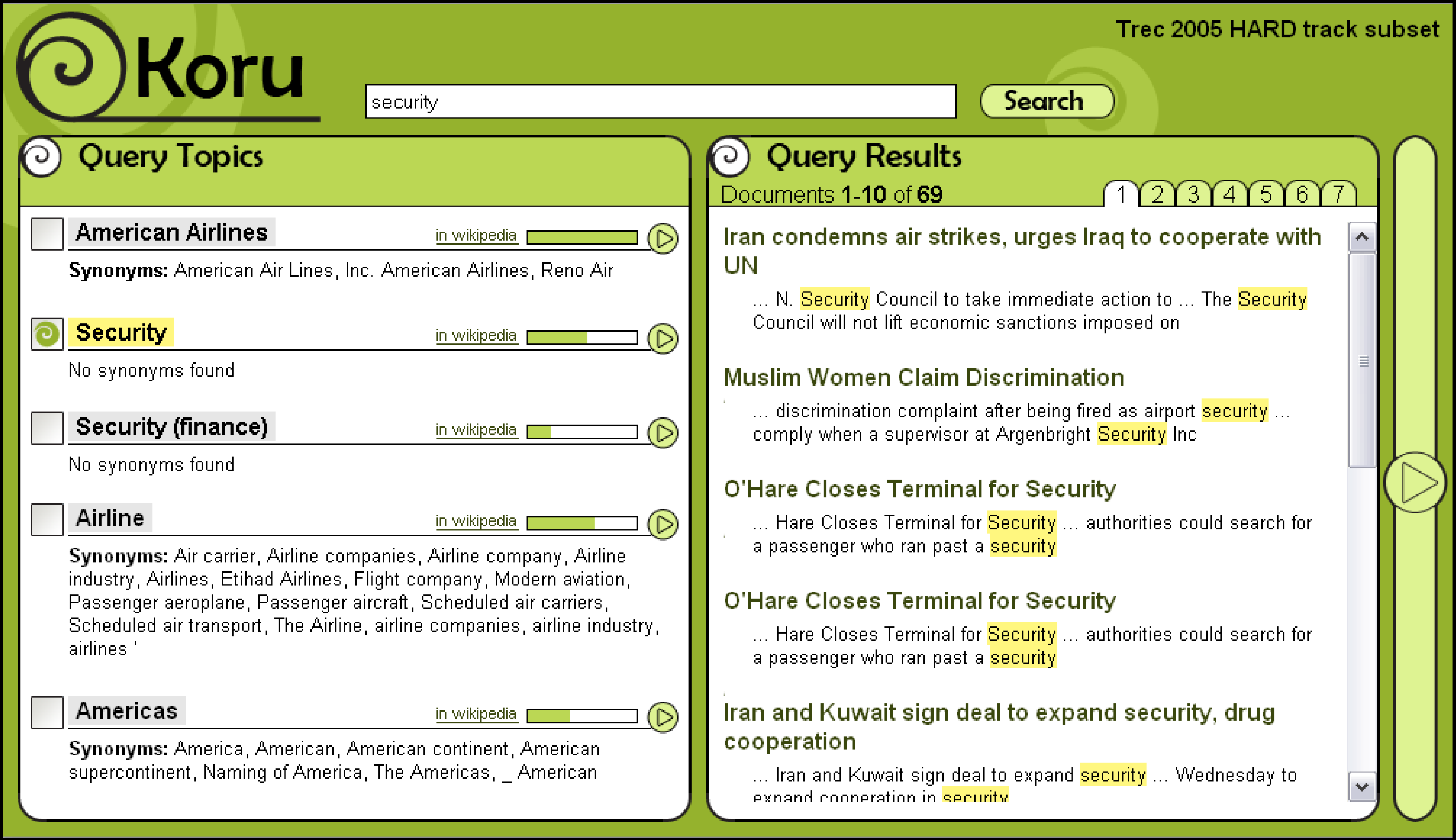
... h1> American Airlines Fined ... 000 civil penalty Thursday against American Airlines because ... their badges in high-security areas at Dallas ... that unless security ... `security identification display area

Airline Workers Held in Drug Probe

... - Security lapses at Miami International Airport and ... s managing director of security ... The drugs were put aboard American Airlines flights in Colombia ... agents in Puerto Rico arrested an American Airlines

Airline Workers Held in Drug Probe

... - Security lapses at Miami International Airport and ... s managing director of security ... The drugs were put aboard American Airlines flights in Colombia ... agents in Puerto Rico arrested an American Airlines



Documents **1-10** of **69**

□ **American Airlines** in wikipedia

Synonyms: American Air Lines, Inc. American Airlines, Reno Air

● **Security** in wikipedia

No synonyms found

□ **Security (finance)** in wikipedia

No synonyms found

□ **Airline** in wikipedia

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

□ **Americas** in wikipedia

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Iran condemns air strikes, urges Iraq to cooperate with UN

... N. Security Council to take immediate action to ... The Security Council will not lift economic sanctions imposed on

Muslim Women Claim Discrimination

... discrimination complaint after being fired as airport security ... comply when a supervisor at Argenbright Security Inc

O'Hare Closes Terminal for Security

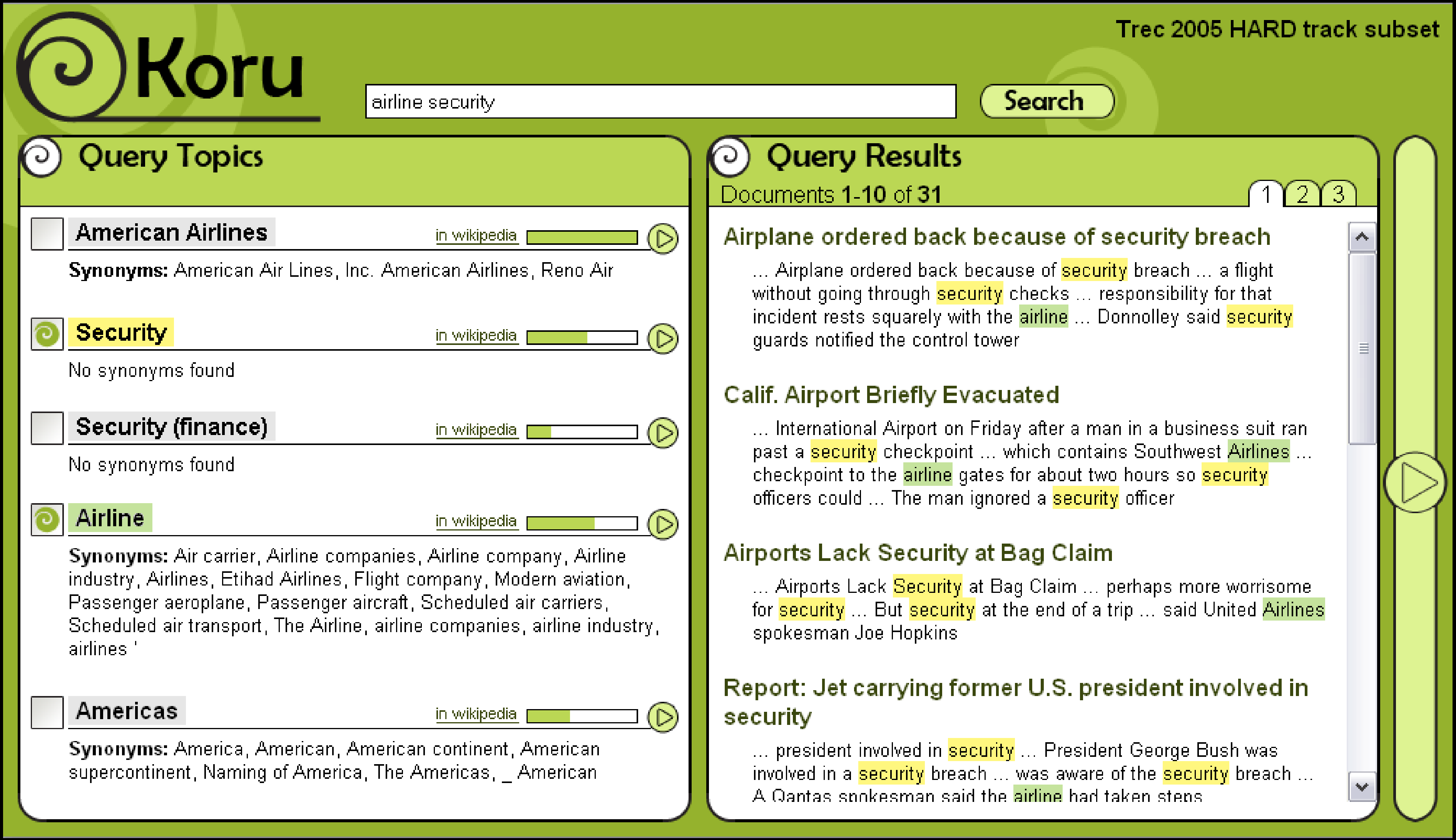
... Hare Closes Terminal for Security ... authorities could search for a passenger who ran past a security

O'Hare Closes Terminal for Security

... Hare Closes Terminal for Security ... authorities could search for a passenger who ran past a security

Iran and Kuwait sign deal to expand security, drug cooperation

... Iran and Kuwait sign deal to expand security ... Wednesday to expand cooperation in security



Documents 1-10 of 31

□ **American Airlines** in wikipedia

Synonyms: American Air Lines, Inc. American Airlines, Reno Air

● **Security** in wikipedia

No synonyms found

□ **Security (finance)** in wikipedia

No synonyms found

● **Airline** in wikipedia

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

□ **Americas** in wikipedia

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Airplane ordered back because of security breach

... Airplane ordered back because of security breach ... a flight without going through security checks ... responsibility for that incident rests squarely with the airline ... Donnolley said security guards notified the control tower

Calif. Airport Briefly Evacuated

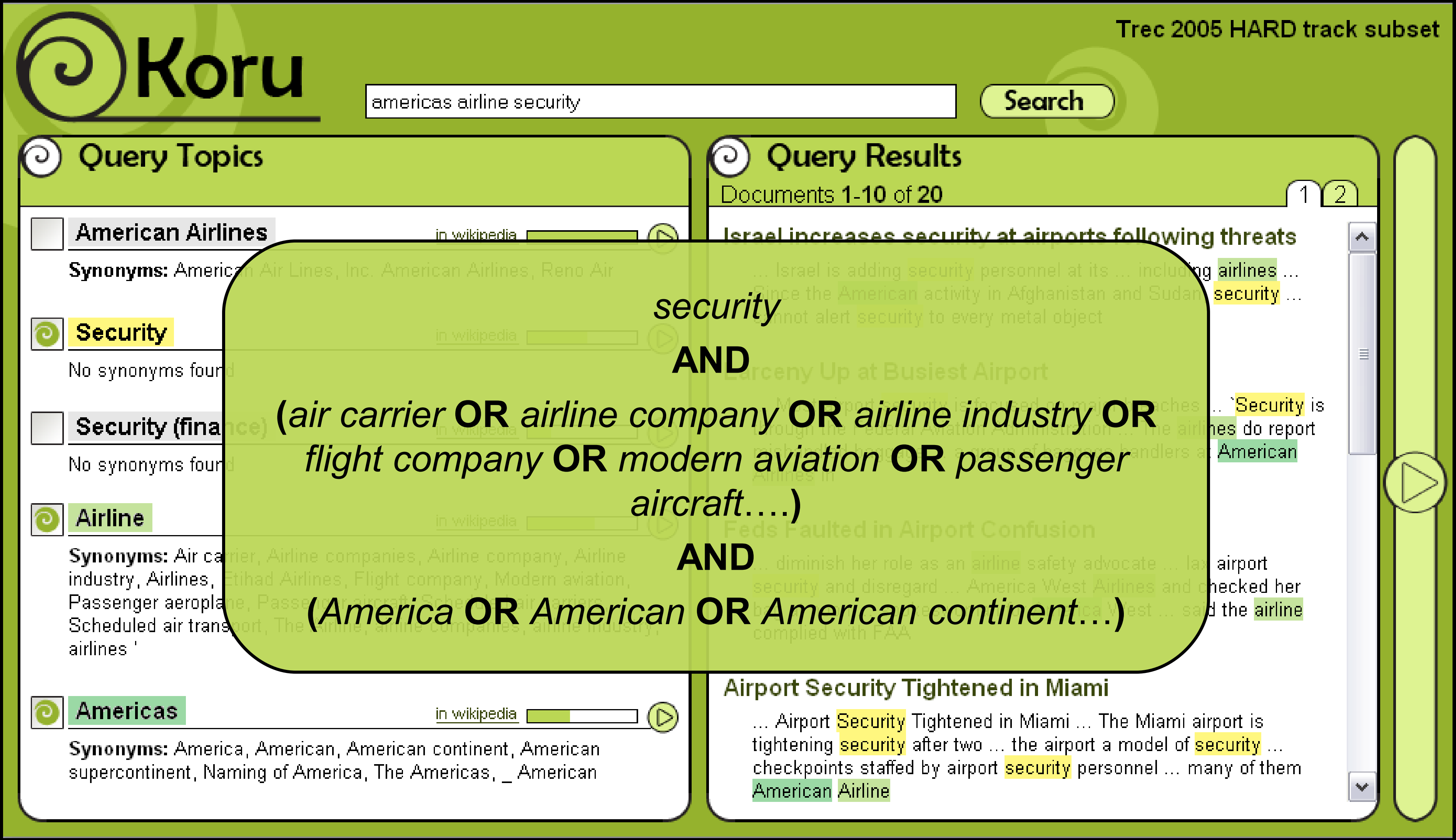
... International Airport on Friday after a man in a business suit ran past a security checkpoint ... which contains Southwest Airlines ... checkpoint to the airline gates for about two hours so security officers could ... The man ignored a security officer

Airports Lack Security at Bag Claim

... Airports Lack Security at Bag Claim ... perhaps more worrisome for security ... But security at the end of a trip ... said United Airlines spokesman Joe Hopkins

Report: Jet carrying former U.S. president involved in security

... president involved in security ... President George Bush was involved in a security breach ... was aware of the security breach ... A Qantas spokesman said the airline had taken steps



Trec 2005 HARD track subset

americas airline security

Query Topics

□ American Airlines

Synonyms: American Air Lines, Inc. American Airlines, Reno Air

Security

No synonyms found

□ Security (finance)

No synonyms found

Airline

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

Americas

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Query Results

Documents 1-10 of 20 1 2

Israel increases security at airports following threats

... Israel is adding security personnel at its ... including airlines ... Since the American activity in Afghanistan and Sudan ... security ... cannot alert security to every metal object

Larceny Up at Busiest Airport

... Miami airport security is focused on major breaches ... `Security is through the Federal Aviation Administration ... The airlines do report mishandled baggage ... a group of baggage handlers at American Airlines in

Feds Faulted in Airport Confusion

... diminish her role as an airline safety advocate ... lax airport security and disregard ... America West Airlines and checked her bag ... America West ... said the airline complied with FAA

Airport Security Tightened in Miami

... Airport Security Tightened in Miami ... The Miami airport is tightening security after two ... the airport a model of security ... checkpoints staffed by airport security personnel ... many of them American Airline

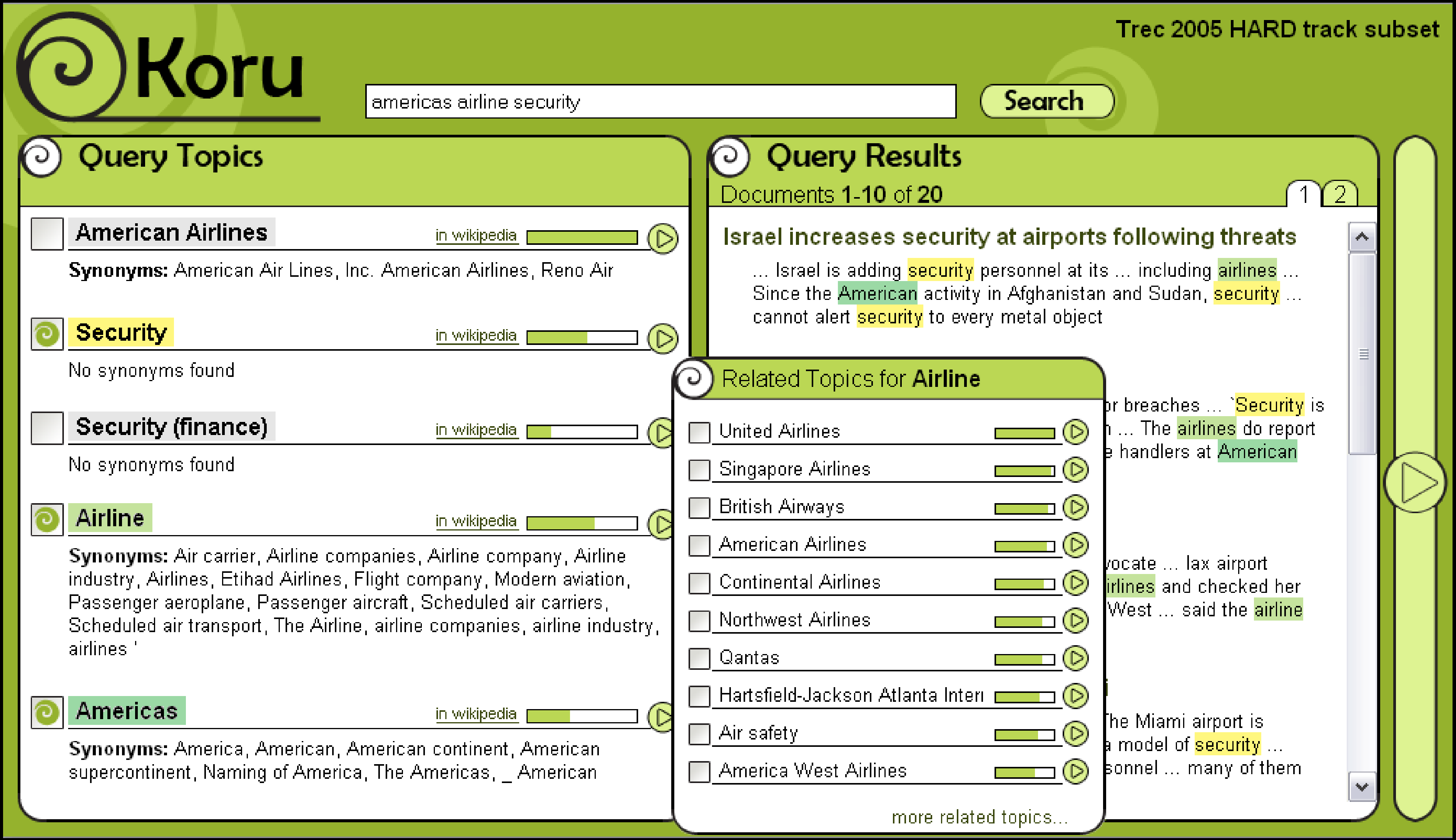
security

AND

(*air carrier* **OR** *airline company* **OR** *airline industry* **OR** *flight company* **OR** *modern aviation* **OR** *passenger aircraft….*)

AND

(*America* **OR** *American* **OR** *American continent…*)



Query Topics

□ American Airlines — in wikipedia

Synonyms: American Air Lines, Inc. American Airlines, Reno Air

Security — in wikipedia

No synonyms found

□ Security (finance) — in wikipedia

No synonyms found

Airline — in wikipedia

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

Americas — in wikipedia

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Query Results

Documents **1-10** of **20** 1 2

Israel increases security at airports following threats

... Israel is adding security personnel at its ... including airlines ... Since the American activity in Afghanistan and Sudan, security ... cannot alert security to every metal object

...r breaches ... `Security is
...n ... The airlines do report
...e handlers at American

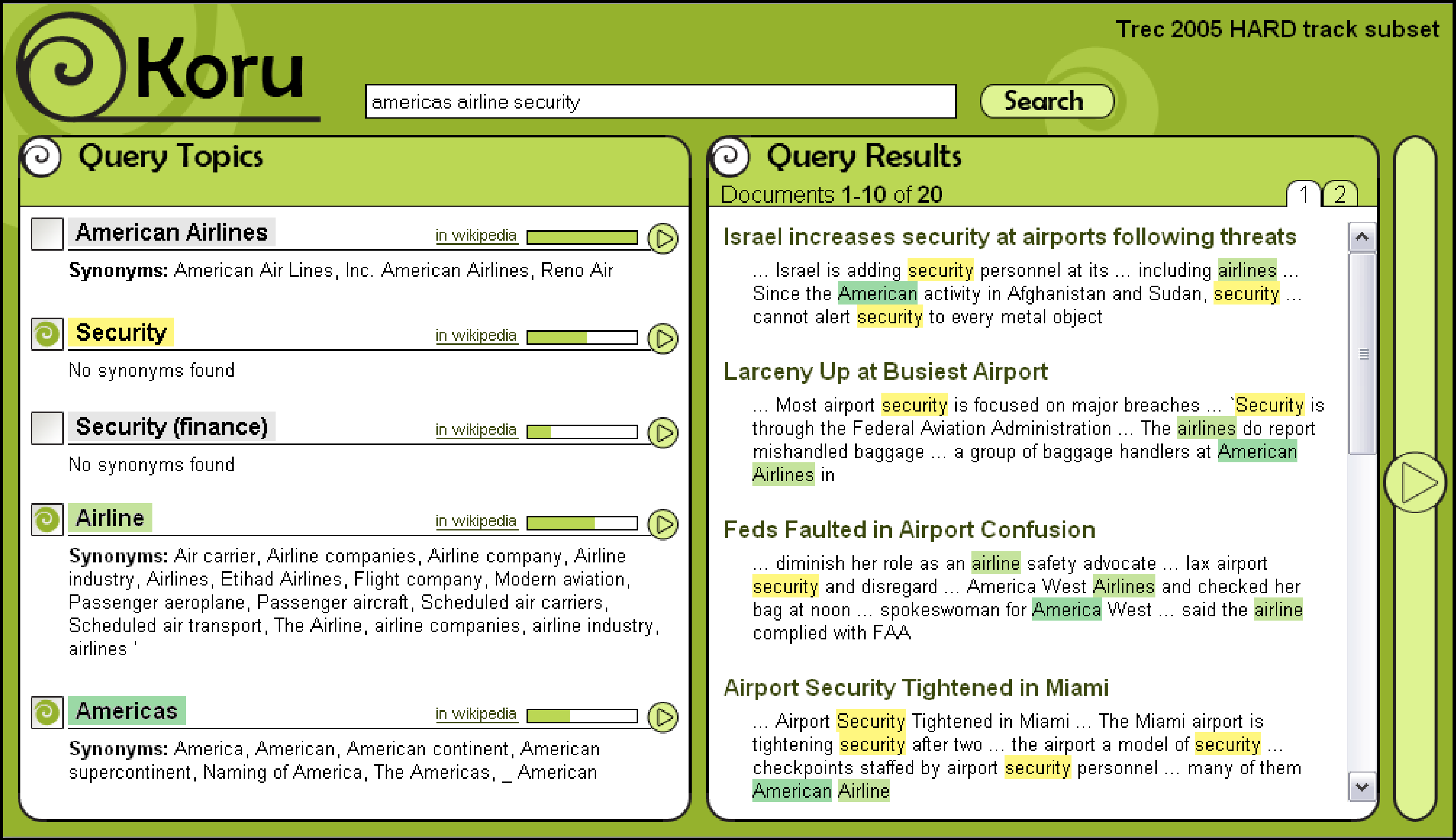
...vocate ... lax airport
...irlines and checked her
...West ... said the airline

...he Miami airport is
...a model of security ...
...sonnel ... many of them

Related Topics for **Airline**

□ United Airlines
□ Singapore Airlines
□ British Airways
□ American Airlines
□ Continental Airlines
□ Northwest Airlines
□ Qantas
□ Hartsfield-Jackson Atlanta Interi
□ Air safety
□ America West Airlines

more related topics...



Trec 2005 HARD track subset

americas airline security

Query Topics

□ **American Airlines** in wikipedia ▷

Synonyms: American Air Lines, Inc. American Airlines, Reno Air

☑ **Security** in wikipedia ▷

No synonyms found

□ **Security (finance)** in wikipedia ▷

No synonyms found

☑ **Airline** in wikipedia ▷

Synonyms: Air carrier, Airline companies, Airline company, Airline industry, Airlines, Etihad Airlines, Flight company, Modern aviation, Passenger aeroplane, Passenger aircraft, Scheduled air carriers, Scheduled air transport, The Airline, airline companies, airline industry, airlines '

☑ **Americas** in wikipedia ▷

Synonyms: America, American, American continent, American supercontinent, Naming of America, The Americas, _ American

Query Results

Documents **1-10** of **20**

1 2

Israel increases security at airports following threats

... Israel is adding security personnel at its ... including airlines ... Since the American activity in Afghanistan and Sudan, security ... cannot alert security to every metal object

Larceny Up at Busiest Airport

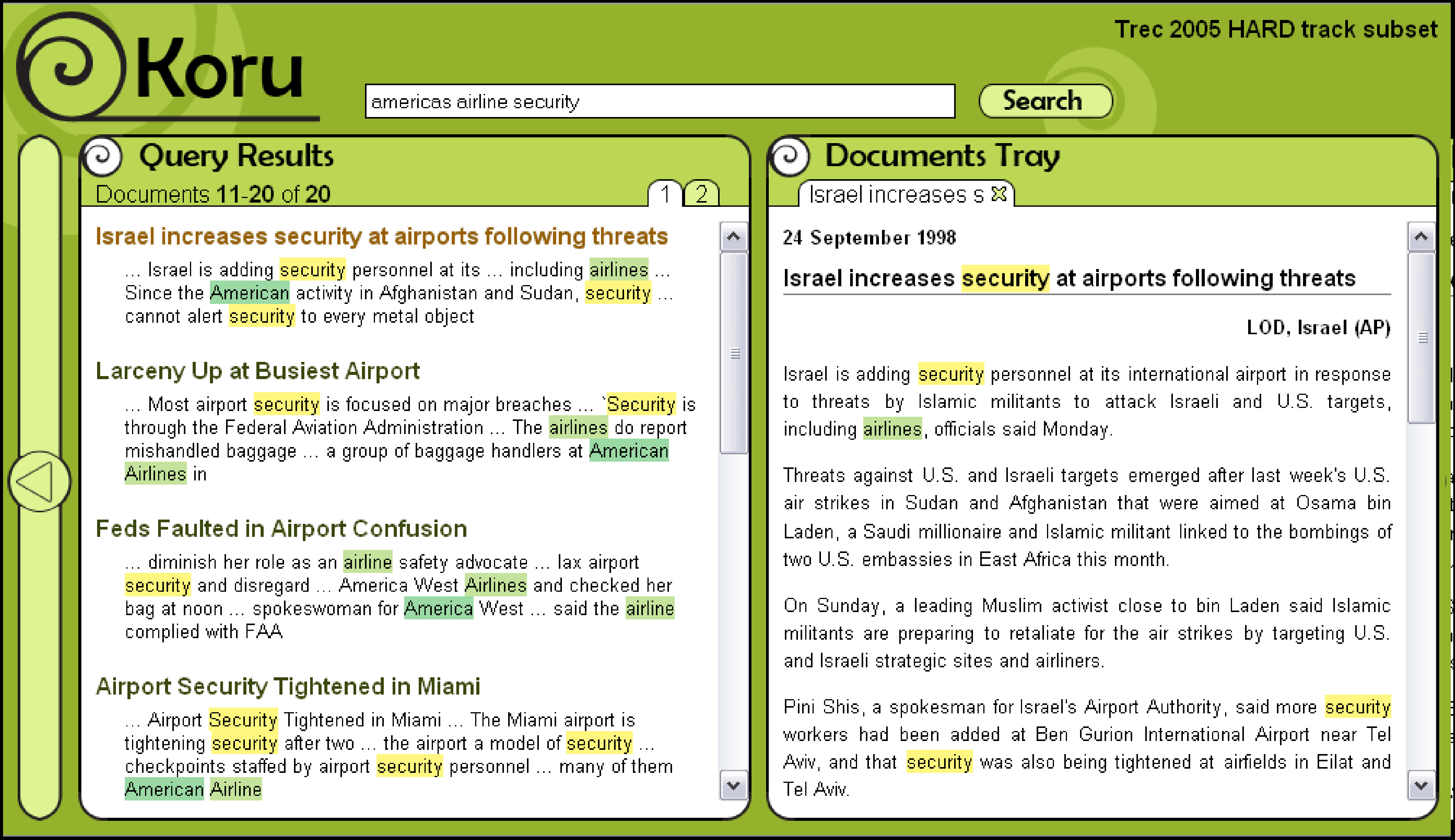
... Most airport security is focused on major breaches ... `Security is through the Federal Aviation Administration ... The airlines do report mishandled baggage ... a group of baggage handlers at American Airlines in

Feds Faulted in Airport Confusion

... diminish her role as an airline safety advocate ... lax airport security and disregard ... America West Airlines and checked her bag at noon ... spokeswoman for America West ... said the airline complied with FAA

Airport Security Tightened in Miami

... Airport Security Tightened in Miami ... The Miami airport is tightening security after two ... the airport a model of security ... checkpoints staffed by airport security personnel ... many of them American Airline



Israel increases security at airports following threats

… Israel is adding security personnel at its … including airlines … Since the American activity in Afghanistan and Sudan, security … cannot alert security to every metal object

Larceny Up at Busiest Airport

… Most airport security is focused on major breaches … `Security is through the Federal Aviation Administration … The airlines do report mishandled baggage … a group of baggage handlers at American Airlines in

Feds Faulted in Airport Confusion

… diminish her role as an airline safety advocate … lax airport security and disregard … America West Airlines and checked her bag at noon … spokeswoman for America West … said the airline complied with FAA

Airport Security Tightened in Miami

… Airport Security Tightened in Miami … The Miami airport is tightening security after two … the airport a model of security … checkpoints staffed by airport security personnel … many of them American Airline

24 September 1998

Israel increases security at airports following threats

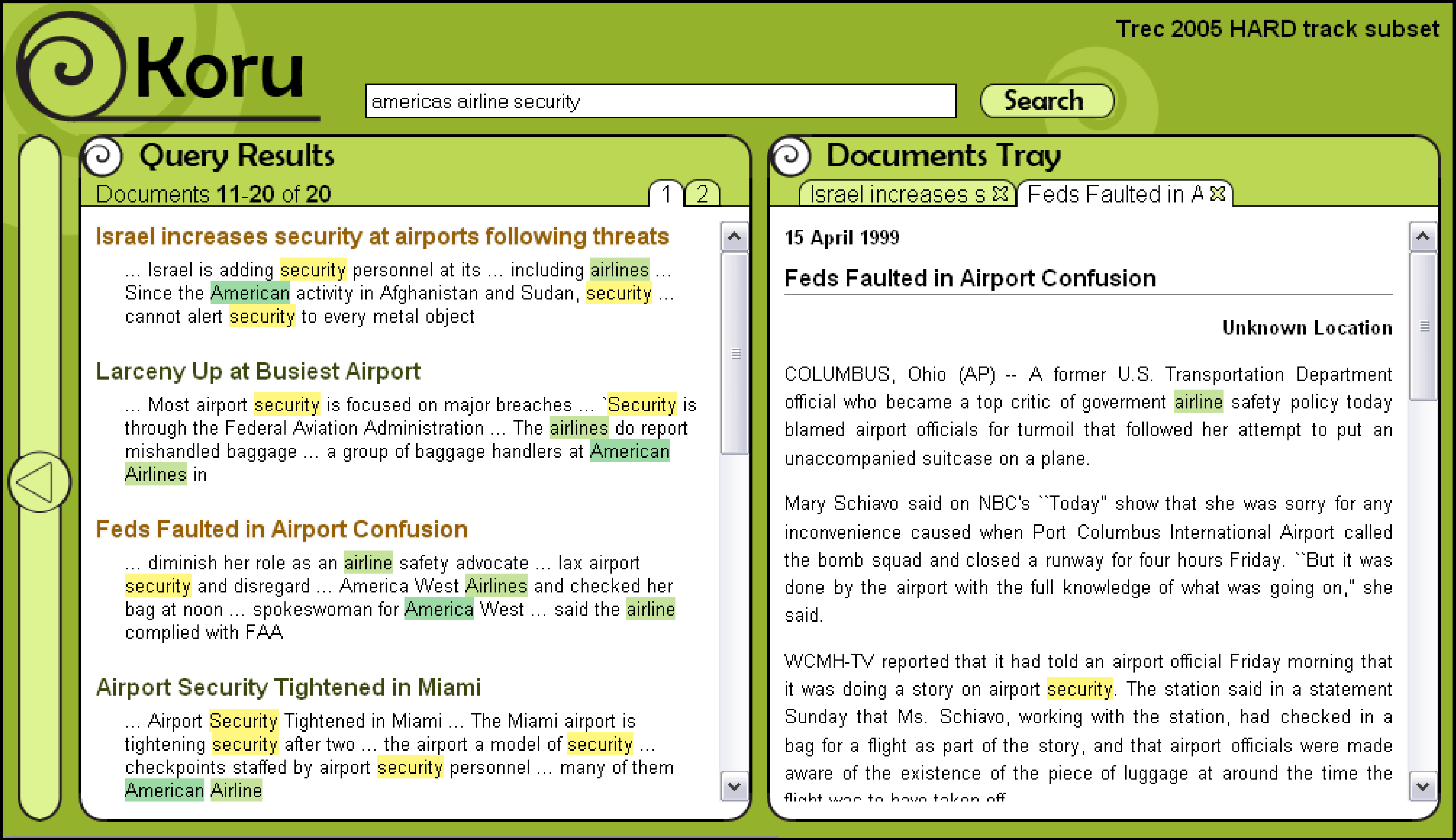
LOD, Israel (AP)

Israel is adding security personnel at its international airport in response to threats by Islamic militants to attack Israeli and U.S. targets, including airlines, officials said Monday.

Threats against U.S. and Israeli targets emerged after last week's U.S. air strikes in Sudan and Afghanistan that were aimed at Osama bin Laden, a Saudi millionaire and Islamic militant linked to the bombings of two U.S. embassies in East Africa this month.

On Sunday, a leading Muslim activist close to bin Laden said Islamic militants are preparing to retaliate for the air strikes by targeting U.S. and Israeli strategic sites and airliners.

Pini Shis, a spokesman for Israel's Airport Authority, said more security workers had been added at Ben Gurion International Airport near Tel Aviv, and that security was also being tightened at airfields in Eilat and Tel Aviv.



Israel increases security at airports following threats

... Israel is adding security personnel at its ... including airlines ... Since the American activity in Afghanistan and Sudan, security ... cannot alert security to every metal object

Larceny Up at Busiest Airport

... Most airport security is focused on major breaches ... `Security is through the Federal Aviation Administration ... The airlines do report mishandled baggage ... a group of baggage handlers at American Airlines in

Feds Faulted in Airport Confusion

... diminish her role as an airline safety advocate ... lax airport security and disregard ... America West Airlines and checked her bag at noon ... spokeswoman for America West ... said the airline complied with FAA

Airport Security Tightened in Miami

... Airport Security Tightened in Miami ... The Miami airport is tightening security after two ... the airport a model of security ... checkpoints staffed by airport security personnel ... many of them American Airline

15 April 1999

Feds Faulted in Airport Confusion

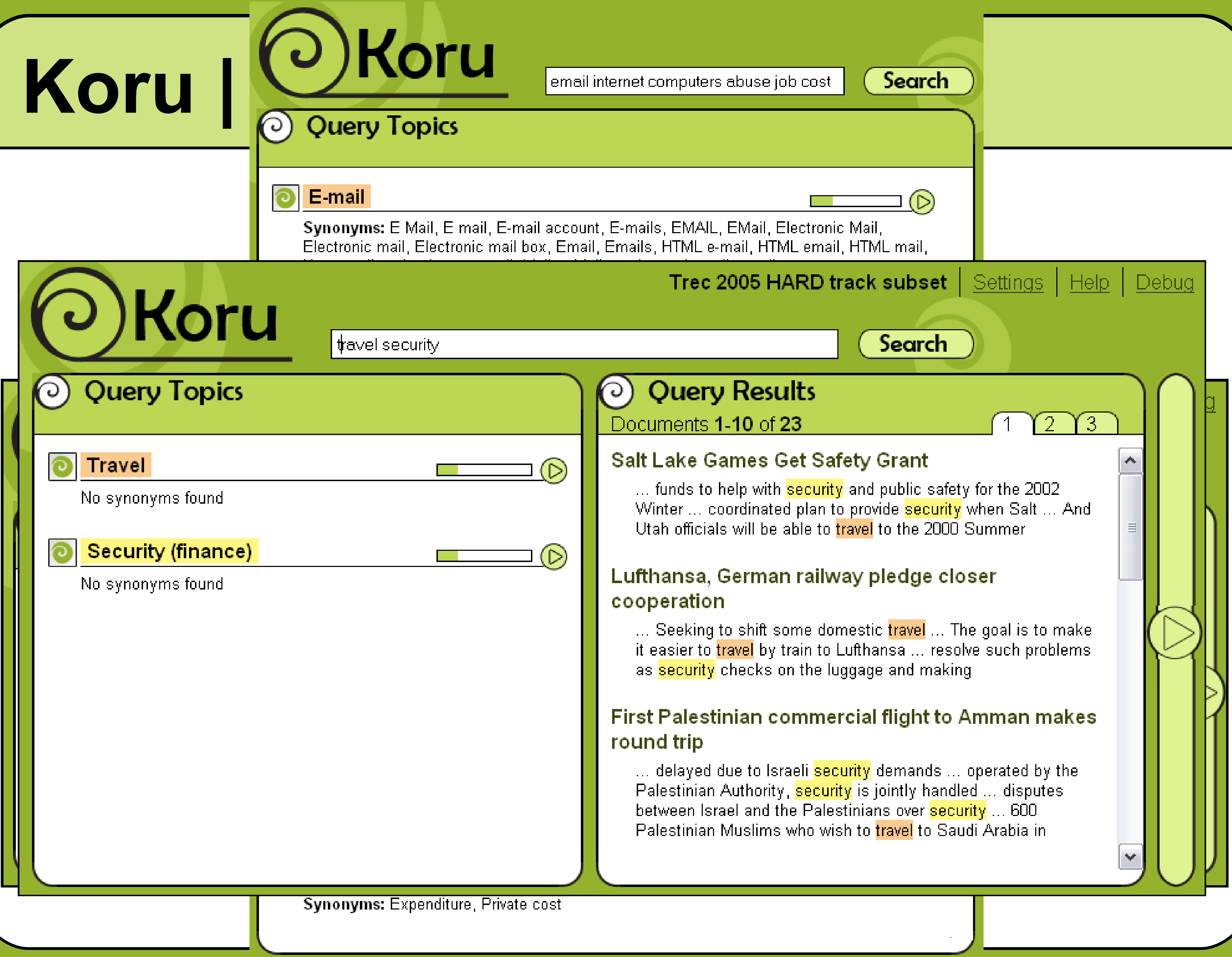
Unknown Location

COLUMBUS, Ohio (AP) -- A former U.S. Transportation Department official who became a top critic of goverment airline safety policy today blamed airport officials for turmoil that followed her attempt to put an unaccompanied suitcase on a plane.

Mary Schiavo said on NBC's ``Today'' show that she was sorry for any inconvenience caused when Port Columbus International Airport called the bomb squad and closed a runway for four hours Friday. ``But it was done by the airport with the full knowledge of what was going on,'' she said.

WCMH-TV reported that it had told an airport official Friday morning that it was doing a story on airport security. The station said in a statement Sunday that Ms. Schiavo, working with the station, had checked in a bag for a flight as part of the story, and that airport officials were made aware of the existence of the piece of luggage at around the time the flight was to have taken off

Koru |



email internet computers abuse job cost

E-mail

Synonyms: E Mail, E mail, E-mail account, E-mails, EMAIL, EMail, Electronic Mail, Electronic mail, Electronic mail box, Email, Emails, HTML e-mail, HTML email, HTML mail,

travel security

Travel

No synonyms found

Security (finance)

No synonyms found

Documents **1-10** of **23**

Salt Lake Games Get Safety Grant

… funds to help with security and public safety for the 2002 Winter … coordinated plan to provide security when Salt … And Utah officials will be able to travel to the 2000 Summer

Lufthansa, German railway pledge closer cooperation

… Seeking to shift some domestic travel … The goal is to make it easier to travel by train to Lufthansa … resolve such problems as security checks on the luggage and making

First Palestinian commercial flight to Amman makes round trip

… delayed due to Israeli security demands … operated by the Palestinian Authority, security is jointly handled … disputes between Israel and the Palestinians over security … 600 Palestinian Muslims who wish to travel to Saudi Arabia in

Synonyms: Expenditure, Private cost

What now

We need to improve how topics and the relations between them are extracted

- Semantic Relatedness
- Wikification

Semantic Relatedness

Given any two terms, what is the strength of the semantic relation between them?

- Highly useful
  - AI, data mining, IR, NLP
- But subjective

Love	Sex	6.77
Tiger	Cat	7.35
Tiger	Tiger	10.00
Book	Paper	7.46
Computer	Keyboard	7.62
Computer	Internet	7.58
Plane	Car	5.77
Stock	Life	0.92
Television	Radio	6.77
…	…	…

Semantic Relatedness | with Wikipedia

Scale and structure

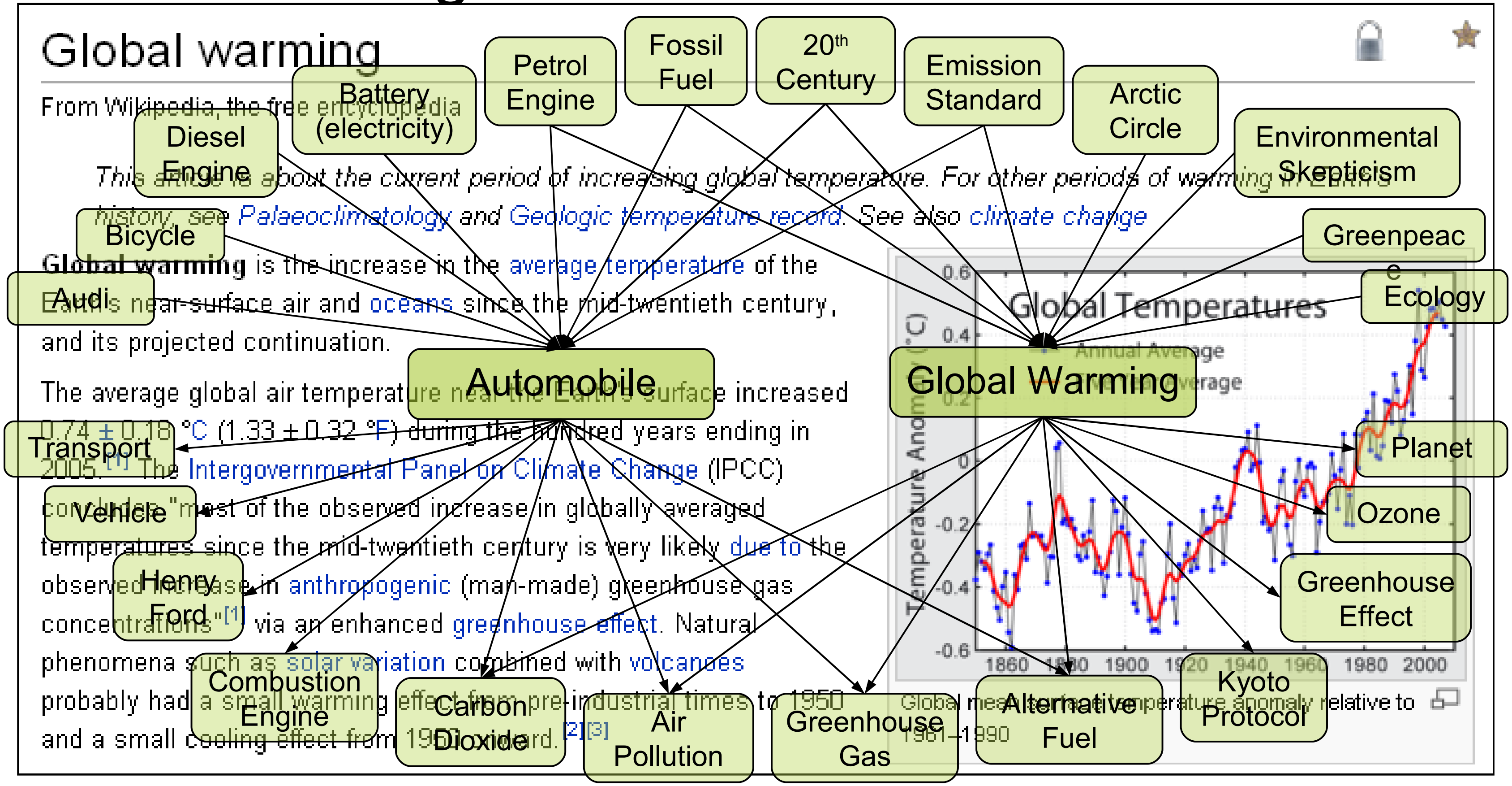
- GBs of text
- millions of articles
- hundreds of thousands of categories

- Two techniques have been developed already

WikiRelate!	19% - 48%
Explicit Semantic Analysis	75%

Semantic Relatedness | Wikipedia links

Wikipedia has an extremely rich hyperlink structure that has been ignored so far.



Global warming

From Wikipedia, the free encyclopedia

This article is about the current period of increasing global temperature. For other periods of warming in Earth's history, see Palaeoclimatology and Geologic temperature record. See also climate change

Global warming is the increase in the average temperature of the Earth's near-surface air and oceans since the mid-twentieth century, and its projected continuation.

The average global air temperature near the Earth's surface increased 0.74 ± 0.18 °C (1.33 ± 0.32 °F) during the hundred years ending in 2005.[1] The Intergovernmental Panel on Climate Change (IPCC) concludes "most of the observed increase in globally averaged temperatures since the mid-twentieth century is very likely due to the observed increase in anthropogenic (man-made) greenhouse gas concentrations"[1] via an enhanced greenhouse effect. Natural phenomena such as solar variation combined with volcanoes probably had a small warming effect from pre-industrial times to 1950 and a small cooling effect from 1950 onward.[2][3]

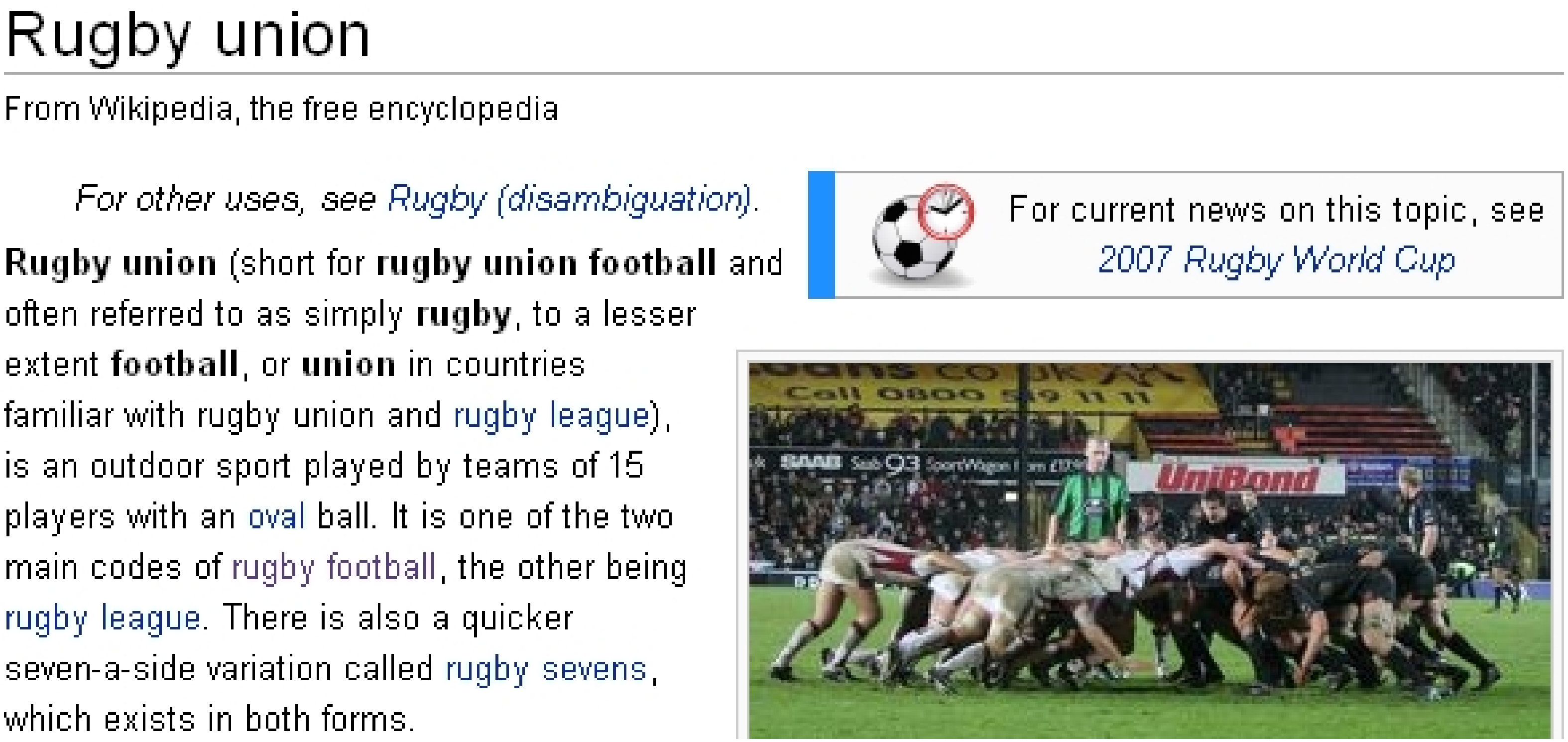
Global mean surface temperature anomaly relative to 1961–1990

Semantic Relatedness | evaluation

Dataset	**WikiRelate**	**ESA**	**WLM**
Miller & Charles	**45%**	**73%**	**70%**
Rubenstein & Goodenough	**52%**	**82%**	**64%**
WordSimilarity 353	**49%**	**75%**	**69%**

WikiRelate **<** WLM **<** ESA

Wikification

- How do we accurately cross-reference documents with Wikipedia?
- 

Rugby union

From Wikipedia, the free encyclopedia

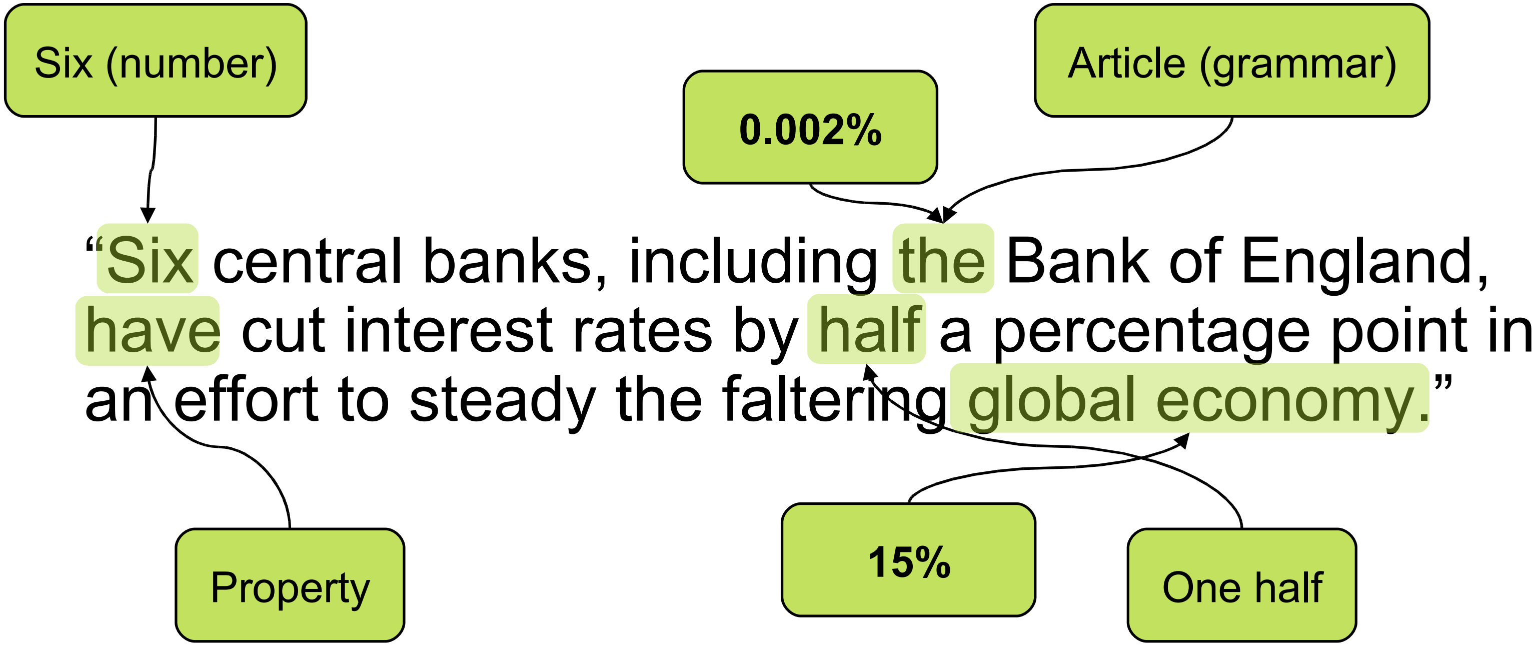
For other uses, see Rugby (disambiguation).

For current news on this topic, see *2007 Rugby World Cup*

Rugby union (short for **rugby union football** and often referred to as simply **rugby**, to a lesser extent **football**, or **union** in countries familiar with rugby union and rugby league), is an outdoor sport played by teams of 15 players with an oval ball. It is one of the two main codes of rugby football, the other being rugby league. There is also a quicker seven-a-side variation called rugby sevens, which exists in both forms.

Wikification | identifying concept terms

- Wikipedia’s links provide a huge vocabulary of which terms can resolve to which concepts



Six (number)

0.002%

Article (grammar)

“Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy.”

Property

15%

One half

Wikification | resolving ambiguity

- For every link in Wikipedia, a human author has manually chosen the correct destination.

"Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy."

Financial institution	97.0%

"The story begins on the banks of the Rio Negro in the Central Amazon. A party of scientists is embarking on a voyage which they hope will provide answers to a five hundred year old mystery."

recall **96%** | precision **98%**

Wikification | selecting relevant concepts

- Wikipedians do not link to every single article
  - only ones that readers would want to investigate

“Six central banks, including the Bank of England, have cut interest rates by half a percentage point in an effort to steady the faltering global economy.”

“The story begins on the banks of the Rio Negro in the Central Amazon. A party of scientists is embarking on a voyage which they hope will provide answers to a five hundred year old mystery.”

recall **74%** **|** precision **74%**

What next?

- Explore applications for Wikification
  - Topic Indexing
  - Document Clustering
  - Document Summarization
- Revisit Koru
  - Apply semantic relatedness and wikification to knowledge base generation, query expansion, and exploratory search
- Write up!

References

Milne, D., Medelyan, O. and Witten, I. H. Mining Domain-Specific Thesauri from Wikipedia: A case study. In *Proceedings of WI 2006*, Hong Kong.

Milne, D., Witten, I.H. and Nichols, D.M. A Knowledge-Based Search Engine Powered by Wikipedia. In *Proceedings of CIKM 2007*, Lisbon, Portugal.

Milne, D. and Witten, I.H. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of WIKIAI 2008*, Chicago, I.L.

Milne, D. and Witten, I.H. Learning to link with Wikipedia. To appear in *Proceedings of CIKM 2008*, Napa Valley, California.

Websites and Demos

www.cs.waikato.ac.nz/~dnk2

www.nzdl.org/koru

wikipedia-miner.sourceforge.net

www.nzdl.org/wikification