CMPE 493
INTRODUCTION TO INFORMATION RETRIEVAL

Introduction

Arzucan Özgür

Department of Computer Engineering, Boğaziçi University

September 29, 2015

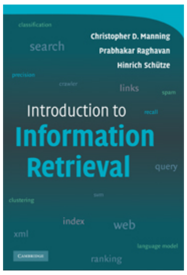
Course Staff

- Instructor: Arzucan Özgür
  - Office: ETA 18
  - Phone: 0212-359-7226
  - E-mail: arzucan.ozgur@boun.edu.tr

  (Please include CMPE493 in your subject when sending e-mail.)
  - Office hours: Monday 14:00-15:00, Tuesday 13:00-15:00, or by appointment.
- TAs:
- Şaziye Betül Özateş (sbetulbilgin@gmail.com)
- Alper Çetiner (alper.cetiner@boun.edu.tr)

Text book

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

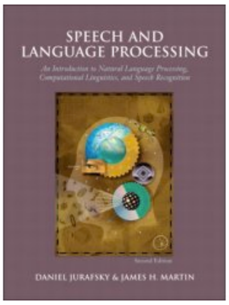


Available online (free) at the website of the book:

http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Reference book (Optional)

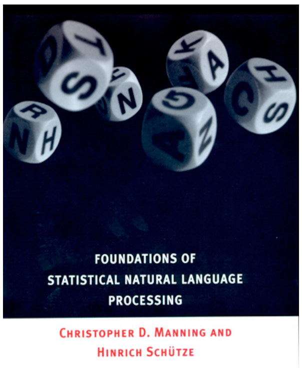
- Daniel Jurafsky and James H. Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, 2008.



Available at the Bookstore.

Reference book (Optional)

- Christopher D. Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
  http://nlp.stanford.edu/fsnlp/



Course Web Site:

- We will use the Moodle Course Management System for lecture notes, announcements, homework/project submissions, and grading.
  - https://moodle.boun.edu.tr

You will automatically be subscribed to the system. You can login using your "boun" e-mail account's username and password.

Grading

- Midterm Exam: 15%
- Final Exam: 15%
- 3-4 Assignments: 30%
- Term Project: 35%
- Class Participation: 5%

Grading – Exams

- In-class midterm and final exams
- Consisting of problems covering the lecture material
- Closed book/notes
- Dates:
  - Midterm Exam: November 4, in the lecture hour (15:00-17:00)
  - Final Exam: As scheduled by the registration office

Grading – Homework Assignments

- Involve some programming where you will implement and test some of the techniques that we cover in class.
- You can use any programming language of your choice such as Perl, Python, Java, etc.
- We should be able to run your program.
- You should provide a readme file, explaining how to run your program.

Term Project

One of the aspects of this course is preparing you for original research in IR.

- Identifying an interesting problem
- Gathering relevant literature and datasets
- Solving it using new algorithms
- Evaluating the results

- Ability to present your ideas and research
  - Writing up your results in a scientific paper format
  - Presenting a research talk to a scientific audience

Term Project

- The project teams can consist of one or two people. (Teams consisting of two people is recommended)
- Each team will choose a project topic by selecting a recent scientific paper from an IR/NLP conference or journal.
- The project will involve replicating the work done in the paper and proposing extensions/improvements to the existing work. The proposed extensions do not need to be implemented.

Some of the Relevant Scientific Conferences

- ACM SIGIR Conference on Research and Development in Information Retrieval
- Conference on Information and Knowledge Management (CIKM)
- ACM International Conference on Web Search and Web Data Mining (WSDM)
- Association for Computational Linguistics (ACL)
- North American Association for Computational Linguistics (NAACL)
- Empirical Methods in Natural Language Processing (EMNLP)
- International Conference on Computational Linguistics (COLING)
- You can select your papers from relevant journals as well, including Information Retrieval, Computational Linguistics, TACL, Natural Language Engineering, and Journal of the Association for Information Science and Technology (JASIST)

Term Project - Deliverables

- Paper selection and 1-2 page description of the methodology planned to be used to replicate the work
  - Date: November 2
- Short project presentation in the end of the semester
  - Tentative Dates: December 15-16, December 22-23, lecture hours
- Submit a project report in the end of the semester.
  - Tentative Date: Final exam date

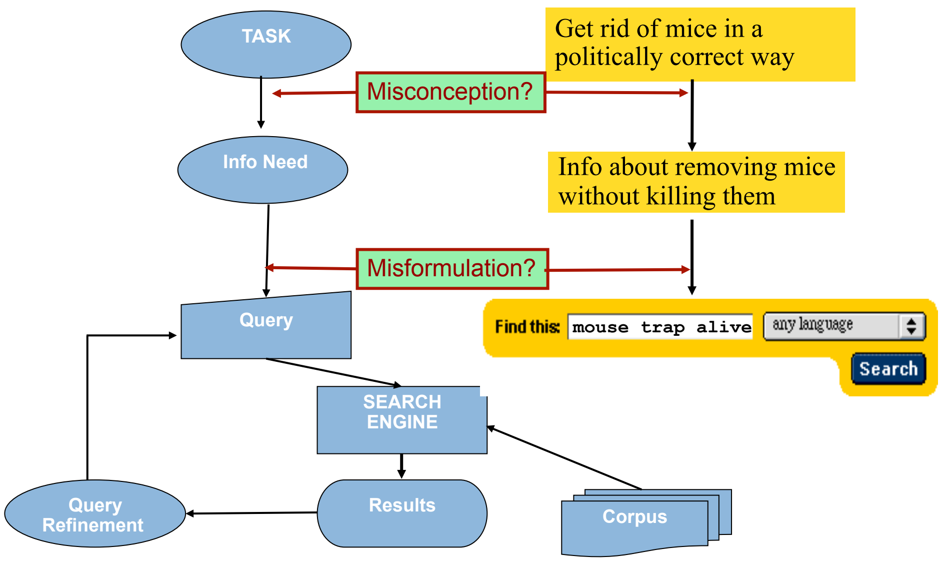
Information Retrieval

- Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections.

Basic assumptions of Information Retrieval

- Collection: Fixed set of documents
- Goal: Retrieve documents with information that is relevant to the user's information need and helps the user complete a task

The classic search model


TASK
Misconception?
Get rid of mice in a politically correct way
Info Need
Info about removing mice without killing them
Misformulation?
Query
Find this: mouse trap alive
any language
Search
SEARCH ENGINE
Results
Corpus
Query Refinement

How good are the retrieved docs?

- *Precision*: Fraction of retrieved docs that are relevant to user's information need
- *Recall*: Fraction of relevant docs in collection that are retrieved

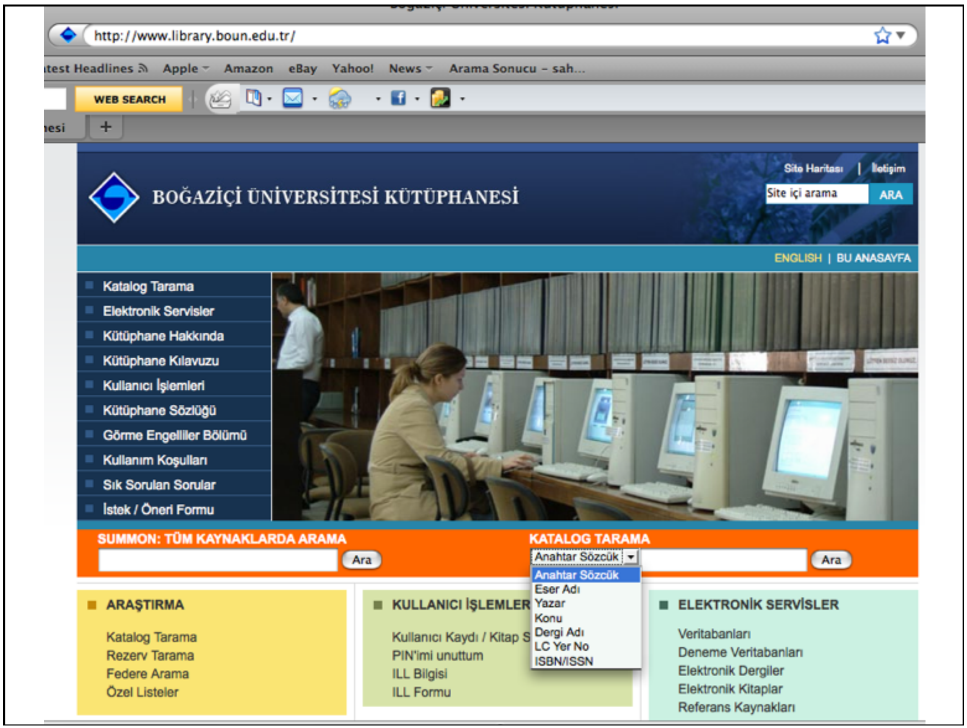
- More precise definitions and measurements to follow in later lectures

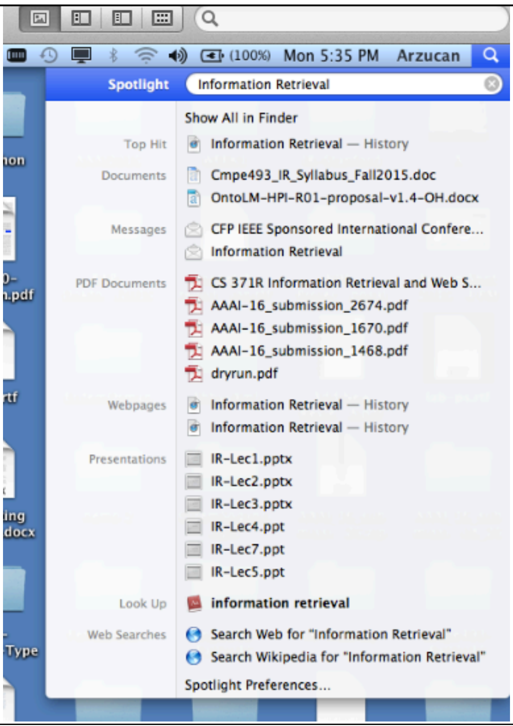
Examples of search engines

- Conventional (library catalog).
  Search by keyword, title, author, etc.
- Text-based (Google, Yahoo!, Bing, Yandex, Baidu; also email search, laptop search etc.)

  Search by keywords. Limited search using queries in natural language.
- Multimedia (QBIC, WebSeek)
  Search by visual appearance (shapes, colors,…).
- Question answering systems (Ask, NSIR, Answerbus)
  Search in (restricted) natural language
- Other:

  music retrieval





Information Retrieval

Show All in Finder

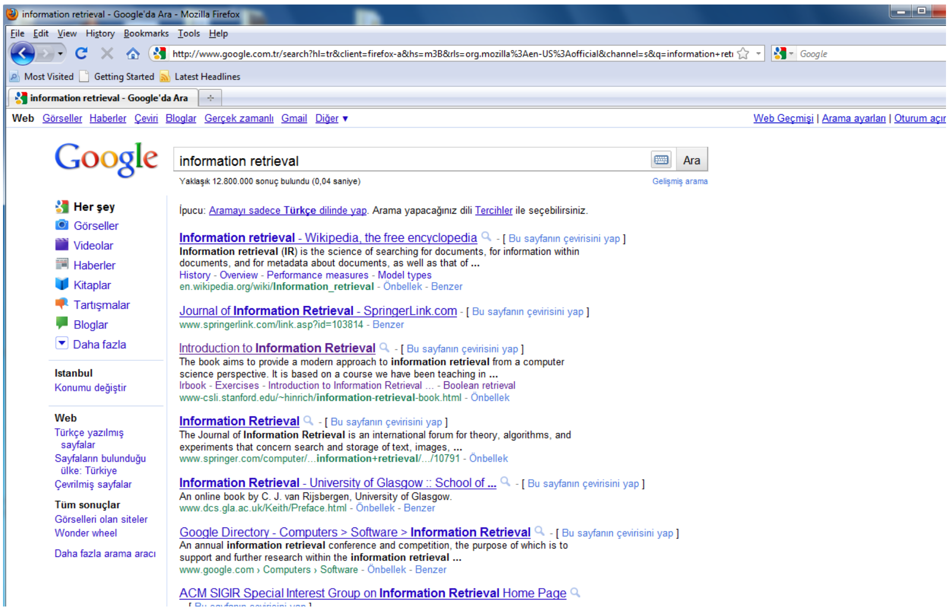
Top Hit	Information Retrieval — History
Documents	Cmpe493_IR_Syllabus_Fall2015.doc OntoLM-HPI-R01-proposal-v1.4-OH.docx
Messages	CFP IEEE Sponsored International Confere... Information Retrieval
PDF Documents	CS 371R Information Retrieval and Web S... AAAI-16_submission_2674.pdf AAAI-16_submission_1670.pdf AAAI-16_submission_1468.pdf dryrun.pdf
Webpages	Information Retrieval — History Information Retrieval — History
Presentations	IR-Lec1.pptx IR-Lec2.pptx IR-Lec3.pptx IR-Lec4.ppt IR-Lec7.ppt IR-Lec5.ppt
Look Up	information retrieval
Web Searches	Search Web for "Information Retrieval" Search Wikipedia for "Information Retrieval"

Spotlight Preferences...

Laptop search
e.g. Spotlight

IR systems on the Web

- Search for Web pages: http://www.google.com
- Domain specific search (e.g., legal, biomedical): PubMed
- Search for images: http://www.picsearch.com
- Search for image content: http://wang14.ist.psu.edu/
- Search for answers to questions: http://www.ask.com
- Music retrieval: http://www.rotorbrain.com/foote/musicr/


information retrieval - Google'da Ara - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com.tr/search?hl=tr&client=firefox-a&hs=m3B&rls=org.mozilla%3Aen-US%3Aofficial&channel=s&q=information+retı
Google
Most Visited Getting Started Latest Headlines
information retrieval - Google'da Ara
Web Görseller Haberler Çeviri Bloglar Gerçek zamanlı Gmail Diğer ▾
Web Geçmişi | Arama ayarları | Oturum açır
Google
information retrieval
Ara
Gelişmiş arama
Her şey
Görseller
Videolar
Haberler
Kitaplar
Tartışmalar
Bloglar
Daha fazla
Istanbul
Konumu değiştir
Web
Türkçe yazılmış sayfalar
Sayfaların bulunduğu ülke: Türkiye
Çevrilmiş sayfalar
Tüm sonuçlar
Görselleri olan siteler
Wonder wheel
Daha fazla arama aracı

Yaklaşık 12.800.000 sonuç bulundu (0,04 saniye)

İpucu: Aramayı sadece **Türkçe** dilinde yap. Arama yapacağınız dili Tercihler ile seçebilirsiniz.

Information retrieval - Wikipedia, the free encyclopedia - [Bu sayfanın çevirisini yap]
Information retrieval (**IR**) is the science of searching for documents, for information within documents, and for metadata about documents, as well as that of **...**
History - Overview - Performance measures - Model types
en.wikipedia.org/wiki/**Information_retrieval** - Önbellek - Benzer

Journal of **Information Retrieval** - SpringerLink.com - [Bu sayfanın çevirisini yap]
www.springerlink.com/link.asp?id=103814 - Benzer

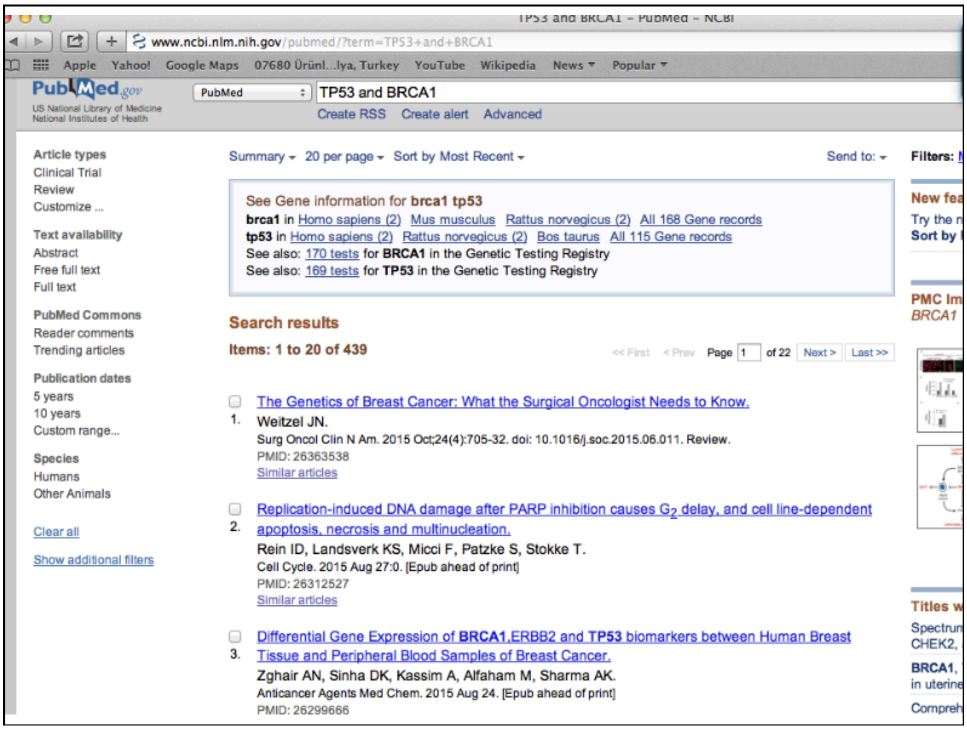
Introduction to **Information Retrieval** - [Bu sayfanın çevirisini yap]
The book aims to provide a modern approach to **information retrieval** from a computer science perspective. It is based on a course we have been teaching in **...**
Irbook - Exercises - Introduction to Information Retrieval ... - Boolean retrieval
www-csli.stanford.edu/~hinrich/**information-retrieval**-book.html - Önbellek

Information Retrieval - [Bu sayfanın çevirisini yap]
The Journal of **Information Retrieval** is an international forum for theory, algorithms, and experiments that concern search and storage of text, images, **...**
www.springer.com/computer/...**information+retrieval**/.../10791 - Önbellek

Information Retrieval - University of Glasgow :: School of **...** - [Bu sayfanın çevirisini yap]
An online book by C. J. van Rijsbergen, University of Glasgow.
www.dcs.gla.ac.uk/Keith/Preface.html - Önbellek - Benzer

Google Directory - Computers > Software > **Information Retrieval** - [Bu sayfanın çevirisini yap]
An annual **information retrieval** conference and competition, the purpose of which is to support and further research within the **information retrieval ...**
www.google.com › Computers › Software - Önbellek - Benzer

ACM SIGIR Special Interest Group on **Information Retrieval** Home Page
[Bu sayfanın çevirisini yap]



TP53 and BRCA1

See Gene information for **brca1 tp53**
brca1 in Homo sapiens (2) Mus musculus Rattus norvegicus (2) All 168 Gene records
tp53 in Homo sapiens (2) Rattus norvegicus (2) Bos taurus All 115 Gene records
See also: 170 tests for **BRCA1** in the Genetic Testing Registry
See also: 169 tests for **TP53** in the Genetic Testing Registry

Search results

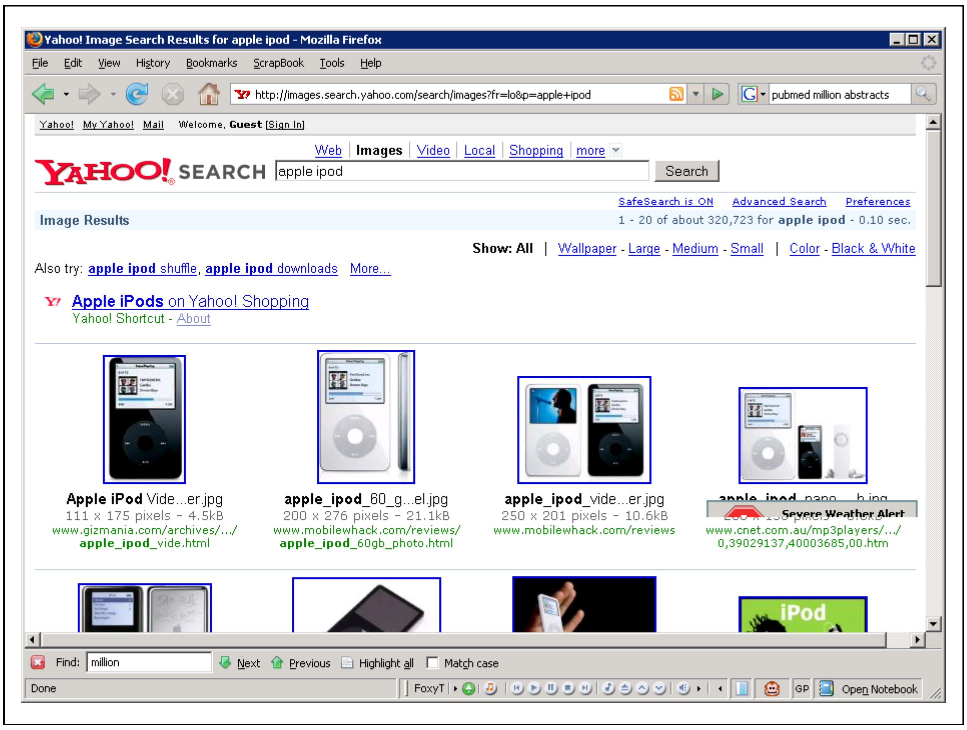
Items: 1 to 20 of 439

<< First < Prev Page 1 of 22 Next > Last >>

□ 1. The Genetics of Breast Cancer: What the Surgical Oncologist Needs to Know.
Weitzel JN.
Surg Oncol Clin N Am. 2015 Oct;24(4):705-32. doi: 10.1016/j.soc.2015.06.011. Review.
PMID: 26363538
Similar articles

□ 2. Replication-induced DNA damage after PARP inhibition causes G_2 delay, and cell line-dependent apoptosis, necrosis and multinucleation.
Rein ID, Landsverk KS, Micci F, Patzke S, Stokke T.
Cell Cycle. 2015 Aug 27:0. [Epub ahead of print]
PMID: 26312527
Similar articles

□ 3. Differential Gene Expression of **BRCA1**,ERBB2 and **TP53** biomarkers between Human Breast Tissue and Peripheral Blood Samples of Breast Cancer.
Zghair AN, Sinha DK, Kassim A, Alfaham M, Sharma AK.
Anticancer Agents Med Chem. 2015 Aug 24. [Epub ahead of print]
PMID: 26299666


Yahoo! Image Search Results for apple ipod - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
http://images.search.yahoo.com/search/images?fr=lo&p=apple+ipod
pubmed million abstracts
Yahoo! My Yahoo! Mail Welcome, Guest [Sign In]
Web Images Video Local Shopping more
YAHOO! SEARCH
Search
SafeSearch is ON Advanced Search Preferences
Severe Weather Alert
iPod
Find: million
Next Previous Highlight all Match case
Done
FoxyT
GP Open Notebook

apple ipod

Image Results

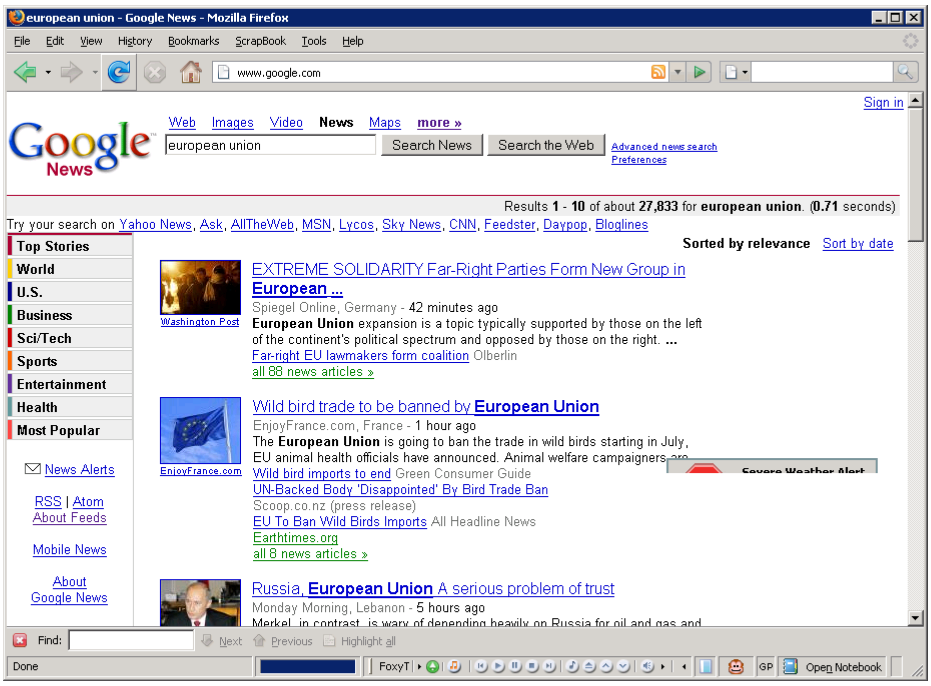
1 - 20 of about 320,723 for **apple ipod** - 0.10 sec.

Show: All | Wallpaper - Large - Medium - Small | Color - Black & White

Also try: **apple ipod** shuffle, **apple ipod** downloads More...

Apple iPods on Yahoo! Shopping
Yahoo! Shortcut - About

Apple iPod Vide...er.jpg 111 x 175 pixels - 4.5kB www.gizmania.com/archives/.../ **apple_ipod**_vide.html	**apple_ipod**_60_g...el.jpg 200 x 276 pixels - 21.1kB www.mobilewhack.com/reviews/ **apple_ipod**_60gb_photo.html	**apple_ipod**_vide...er.jpg 250 x 201 pixels - 10.6kB www.mobilewhack.com/reviews	**apple_ipod**_nano_...h.jpg www.cnet.com.au/mp3players/.../ 0,39029137,40003685,00.htm



european union - Google News - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
www.google.com
Sign in
Web Images Video News Maps more »
Search News
Search the Web
Advanced news search
Preferences
Find:
Next
Previous
Highlight all
Done
FoxyT
GP
Open Notebook

european union

Results **1** - **10** of about **27,833** for **european union**. (**0.71** seconds)

Try your search on Yahoo News, Ask, AllTheWeb, MSN, Lycos, Sky News, CNN, Feedster, Daypop, Bloglines

Sorted by relevance Sort by date

- Top Stories
- World
- U.S.
- Business
- Sci/Tech
- Sports
- Entertainment
- Health
- Most Popular

News Alerts

RSS | Atom
About Feeds

Mobile News

About
Google News

Washington Post

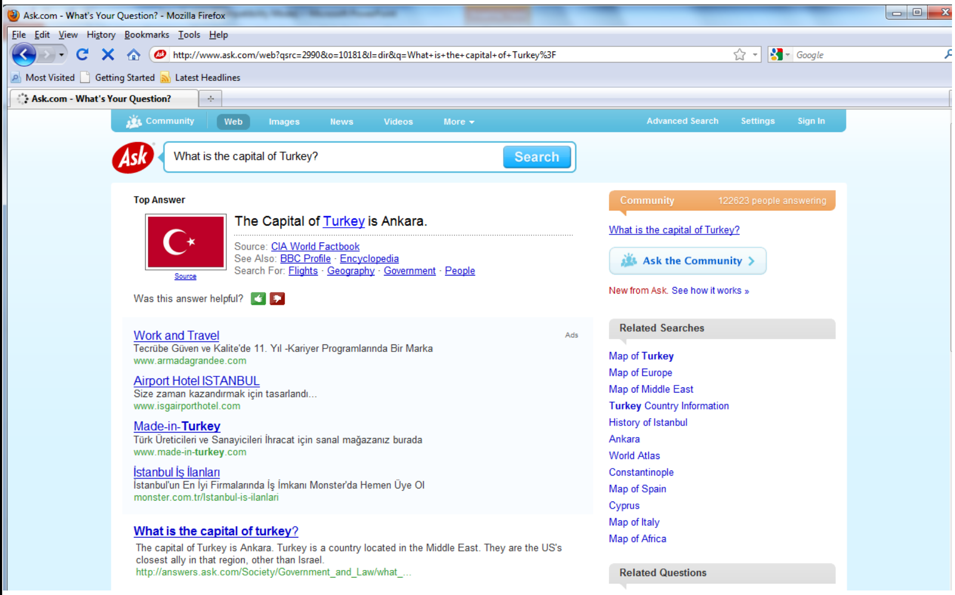
EXTREME SOLIDARITY Far-Right Parties Form New Group in **European ...**
Spiegel Online, Germany - 42 minutes ago
European Union expansion is a topic typically supported by those on the left of the continent's political spectrum and opposed by those on the right. **...**
Far-right EU lawmakers form coalition Olberlin
all 88 news articles »

EnjoyFrance.com

Wild bird trade to be banned by **European Union**
EnjoyFrance.com, France - 1 hour ago
The **European Union** is going to ban the trade in wild birds starting in July, EU animal health officials have announced. Animal welfare campaigners are
Wild bird imports to end Green Consumer Guide
UN-Backed Body 'Disappointed' By Bird Trade Ban
Scoop.co.nz (press release)
EU To Ban Wild Birds Imports All Headline News
Earthtimes.org
all 8 news articles »

Severe Weather Alert

Russia, **European Union** A serious problem of trust
Monday Morning, Lebanon - 5 hours ago
Merkel, in contrast, is wary of depending heavily on Russia for oil and gas and



http://www.ask.com/web?qsrc=2990&o=10181&l=dir&q=What+is+the+capital+of+Turkey%3F

What is the capital of Turkey?

Top Answer

The Capital of Turkey is Ankara.

Source

Source: CIA World Factbook
See Also: BBC Profile · Encyclopedia
Search For: Flights · Geography · Government · People

Was this answer helpful?

Ads

Work and Travel
Tecrübe Güven ve Kalite'de 11. Yıl -Kariyer Programlarında Bir Marka
www.armadagrandee.com

Airport Hotel ISTANBUL
Size zaman kazandırmak için tasarlandı...
www.isgairporthotel.com

Made-in-**Turkey**
Türk Üreticileri ve Sanayicileri İhracat için sanal mağazanız burada
www.made-in-**turkey**.com

İstanbul İş İlanları
İstanbul'un En İyi Firmalarında İş İmkanı Monster'da Hemen Üye Ol
monster.com.tr/Istanbul-is-ilanlari

What is the capital of turkey?
The capital of Turkey is Ankara. Turkey is a country located in the Middle East. They are the US's closest ally in that region, other than Israel.
http://answers.ask.com/Society/Government_and_Law/what_...

Community 122623 people answering

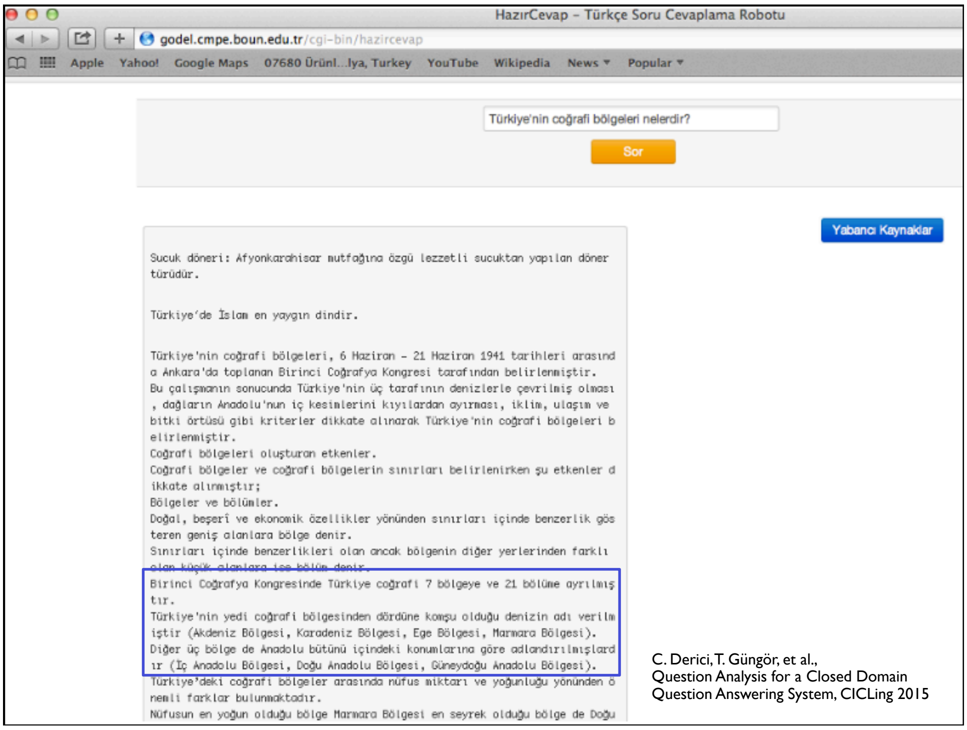
What is the capital of Turkey?

New from Ask. See how it works »

Related Searches

- Map of **Turkey**
- Map of Europe
- Map of Middle East
- **Turkey** Country Information
- History of Istanbul
- Ankara
- World Atlas
- Constantinople
- Map of Spain
- Cyprus
- Map of Italy
- Map of Africa

Related Questions



HazırCevap – Türkçe Soru Cevaplama Robotu
godel.cmpe.boun.edu.tr/cgi-bin/hazircevap
Apple
Yahoo!
Google Maps
07680 Ürünl...lya, Turkey
YouTube
Wikipedia
News
Popular
Sor
Yabancı Kaynaklar

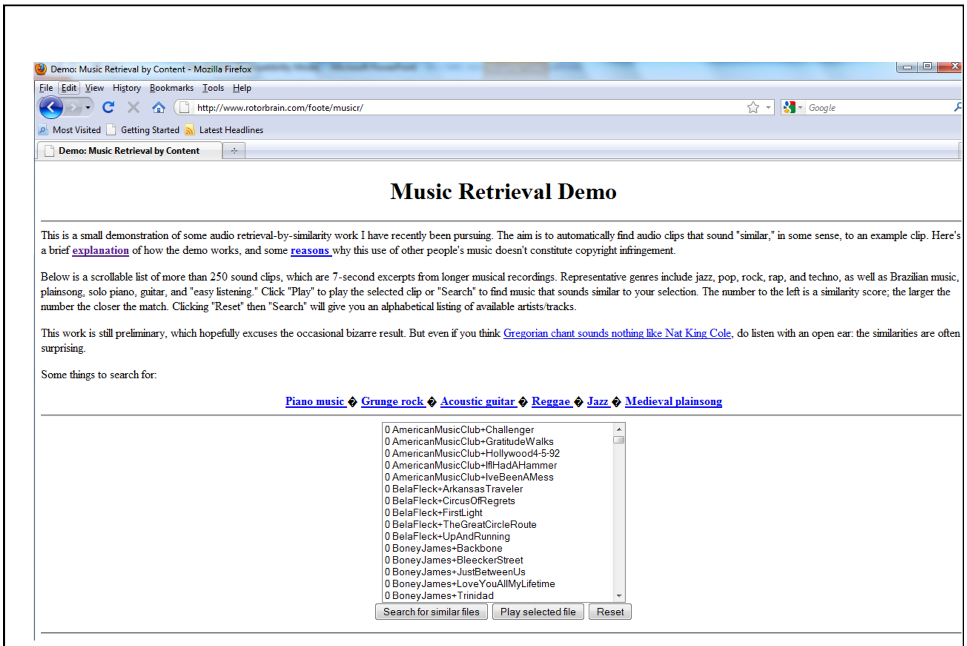
Türkiye'nin coğrafi bölgeleri nelerdir?

```
Sucuk döneri: Afyonkarahisar mutfağına özgü lezzetli sucuktan yapılan döner
türüdür.


Türkiye’de İslam en yaygın dindir.


Türkiye'nin coğrafi bölgeleri, 6 Haziran – 21 Haziran 1941 tarihleri arasınd
a Ankara'da toplanan Birinci Coğrafya Kongresi tarafından belirlenmiştir.
Bu çalışmanın sonucunda Türkiye'nin üç tarafının denizlerle çevrilmiş olması
, dağların Anadolu'nun iç kesimlerini kıyılardan ayırması, iklim, ulaşım ve
bitki örtüsü gibi kriterler dikkate alınarak Türkiye'nin coğrafi bölgeleri b
elirlenmiştir.
Coğrafi bölgeleri oluşturan etkenler.
Coğrafi bölgeler ve coğrafi bölgelerin sınırları belirlenirken şu etkenler d
ikkate alınmıştır;
Bölgeler ve bölümler.
Doğal, beşerî ve ekonomik özellikler yönünden sınırları içinde benzerlik gös
teren geniş alanlara bölge denir.
Sınırları içinde benzerlikleri olan ancak bölgenin diğer yerlerinden farklı
olan küçük alanlara ise bölüm denir.
Birinci Coğrafya Kongresinde Türkiye coğrafi 7 bölgeye ve 21 bölüme ayrılmış
tır.
Türkiye'nin yedi coğrafi bölgesinden dördüne komşu olduğu denizin adı verilm
iştir (Akdeniz Bölgesi, Karadeniz Bölgesi, Ege Bölgesi, Marmara Bölgesi).
Diğer üç bölge de Anadolu bütünü içindeki konumlarına göre adlandırılmışlard
ır (İç Anadolu Bölgesi, Doğu Anadolu Bölgesi, Güneydoğu Anadolu Bölgesi).
Türkiye’deki coğrafi bölgeler arasında nüfus miktarı ve yoğunluğu yönünden ö
nemli farklar bulunmaktadır.
Nüfusun en yoğun olduğu bölge Marmara Bölgesi en seyrek olduğu bölge de Doğu
```

C. Derici, T. Güngör, et al.,
Question Analysis for a Closed Domain
Question Answering System, CICLing 2015



Demo: Music Retrieval by Content - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.rotorbrain.com/foote/musicr/
Google
Most Visited
Getting Started
Latest Headlines
Demo: Music Retrieval by Content

Music Retrieval Demo

This is a small demonstration of some audio retrieval-by-similarity work I have recently been pursuing. The aim is to automatically find audio clips that sound "similar," in some sense, to an example clip. Here's a brief explanation of how the demo works, and some reasons why this use of other people's music doesn't constitute copyright infringement.

Below is a scrollable list of more than 250 sound clips, which are 7-second excerpts from longer musical recordings. Representative genres include jazz, pop, rock, rap, and techno, as well as Brazilian music, plainsong, solo piano, guitar, and "easy listening." Click "Play" to play the selected clip or "Search" to find music that sounds similar to your selection. The number to the left is a similarity score; the larger the number the closer the match. Clicking "Reset" then "Search" will give you an alphabetical listing of available artists/tracks.

This work is still preliminary, which hopefully excuses the occasional bizarre result. But even if you think Gregorian chant sounds nothing like Nat King Cole, do listen with an open ear: the similarities are often surprising.

Some things to search for:

Piano music � Grunge rock � Acoustic guitar � Reggae � Jazz � Medieval plainsong

```
0 AmericanMusicClub+Challenger
0 AmericanMusicClub+GratitudeWalks
0 AmericanMusicClub+Hollywood4-5-92
0 AmericanMusicClub+IfIHadAHammer
0 AmericanMusicClub+IveBeenAMess
0 BelaFleck+ArkansasTraveler
0 BelaFleck+CircusOfRegrets
0 BelaFleck+FirstLight
0 BelaFleck+TheGreatCircleRoute
0 BelaFleck+UpAndRunning
0 BoneyJames+Backbone
0 BoneyJames+BleeckerStreet
0 BoneyJames+JustBetweenUs
0 BoneyJames+LoveYouAllMyLifetime
0 BoneyJames+Trinidad
```

What does it take to build a search engine?

- Decide what to index
- Collect it
- Index it (efficiently)
- Keep the index up to date
- Provide user-friendly query facilities

What else?

- Understand the structure of the web for efficient crawling
- Understand user information needs
- Preprocess unstructured textual data
- Cluster data
- Classify data
- Evaluate performance

Goals of the course

- Understand how search engines work
- Understand the limits of existing search technology
- Learn about the state of the art in IR research
- Learn to analyze textual data sets
- Learn to evaluate information retrieval systems
- Learn about standardized document collections
- Learn about text similarity measures
- Learn about semantic dimensionality reduction
- Learn about web crawling
- Learn to use existing software
- Understand the dynamics of the Web by building appropriate mathematical models
- Build working systems that assist users in finding useful information from large collections

Topics (tentative list)

- Boolean model; text pre-processing; inverted indexes
- Approximate string matching and tolerant retrieval
- Index construction and compression
- Vector space model; text-similarity metrics; term weighting; ranked retrieval
- Evaluating information retrieval systems
- Relevance feedback; query expansion
- Language models for information retrieval
- Text classification and clustering
- Latent semantic indexing
- Web search and crawling
- Link analysis (e.g. hubs and authorities, Google PageRank)

References

- Content adapted from Prof. Dragomir Radev and the IR book's web site.