CMPE 493
INTRODUCTION TO INFORMATION RETRIEVAL

Introduction

Arzucan Özgür

Department of Computer Engineering, Boğaziçi University
October 26, 2020

Course Staff

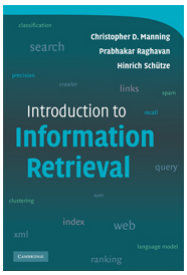
- Instructor: Arzucan Özgür
  - Office: BM 18
  - Phone: 0212-359-7226
  - E-mail: arzucan.ozgur@boun.edu.tr

(Please include CMPE493 in your subject when sending e-mail.)

- Office hours: Monday after class or by appointment. Please send me an e-mail in advance.

Text book

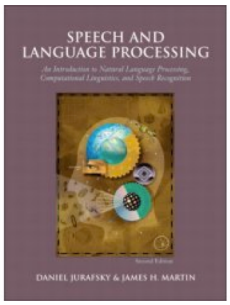
- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.



Available at the bookstore and the library.
Also, available online (free) at the website of the book:
http://nlp.stanford.edu/IR-book/information-retrieval-book.html

Reference book (Optional)

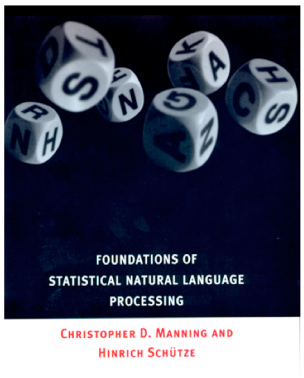
- Daniel Jurafsky and James H. Martin, SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition, 2008.



Available at the Bookstore and the library. Draft of the 3rd edition available at:
https://web.stanford.edu/~jurafsky/slp3/

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

- 4 or 5 Programming Assignments: 55%
- 2 Problem Sets: 10%
- Term Project: 35%

Grading – Programming Assignments

- Involve intermediate-level 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 (Python, Java, C++ etc.).
- We should be able to run your program.
- You should provide a readme file, explaining how to run your program.

Term Project

- Project teams will consist of two or three people.
- We will have a shared task, where all teams work on the same problem.
- Deliverables:
  - December 7th, Monday (lecture hour): Each team will give a 10min presentation in front of the class describing their progress so far.
  - In the final exam slot: Each team will give a 15min presentation and demo in front of the class describing their system and results.

Term Project

- Annotated data will be provided. The project will involve the remaining steps of designing and developing a useful text processing/analysis system including:
  - System design
  - System development
  - System evaluation

Some of the Relevant Scientific Conferences

- ACM SIGIR Conference on Research and Development in Information Retrieval
- Shared Tasks such SemEval and BioNLP
- 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)
- Some relevant journals: Information Retrieval, Computational Linguistics, TACL, Natural Language Engineering, Journal of the Association for Information Science and Technology (JASIST)

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.

IR vs. databases: Structured vs unstructured data

Employee	Manager	Salary
Smith	Jones	50000
Chang	Smith	60000
Ivy	Smith	50000

Typically allows numerical range and exact match (or pattern match using the LIKE operator) (for text) queries, e.g.,
Salary < 60000 AND Manager = Smith.

Unstructured data

- Typically refers to free text
- Allows
  - Keyword queries including operators
  - More sophisticated "concept" queries e.g.,
    - find all web pages about *infectious diseases*

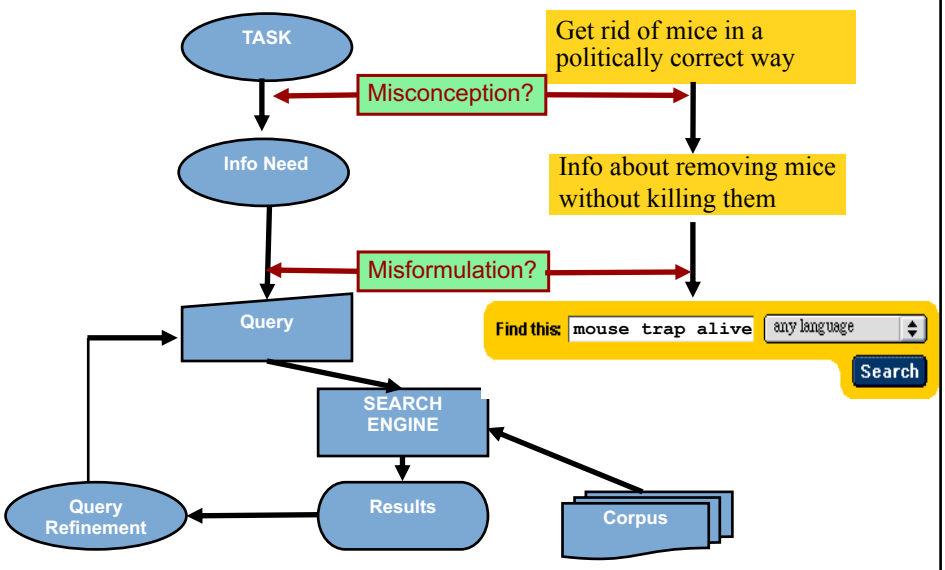
Semi-structured data

- In fact almost no data is "unstructured"
- E.g., this slide has distinctly identified zones such as the *Title* and *Bullets*
- Facilitates "semi-structured" search such as
  - *Title* contains data AND *Bullets* contain search

Basic assumptions of Information Retrieval

- Collection: 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

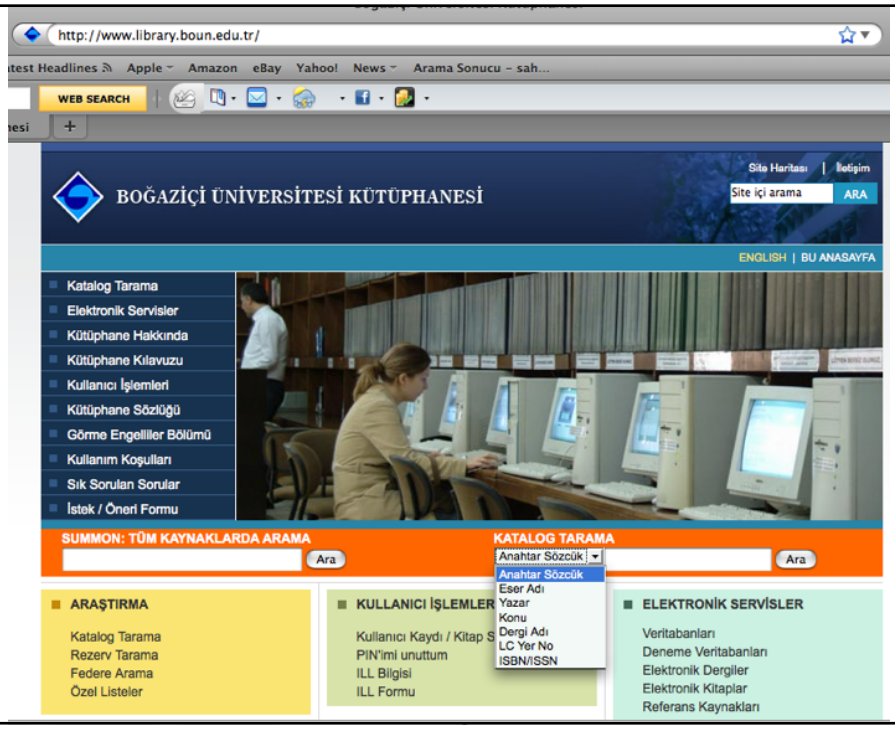


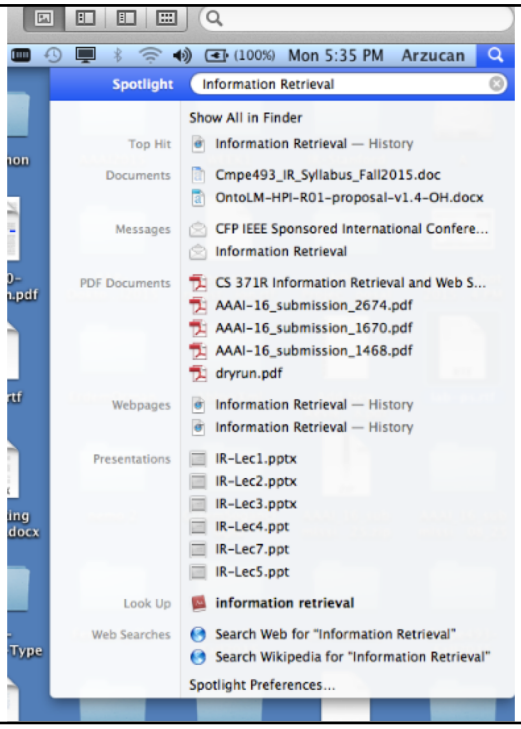
How good are the retrieved docs?

- *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 (DuckDuckGo, 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

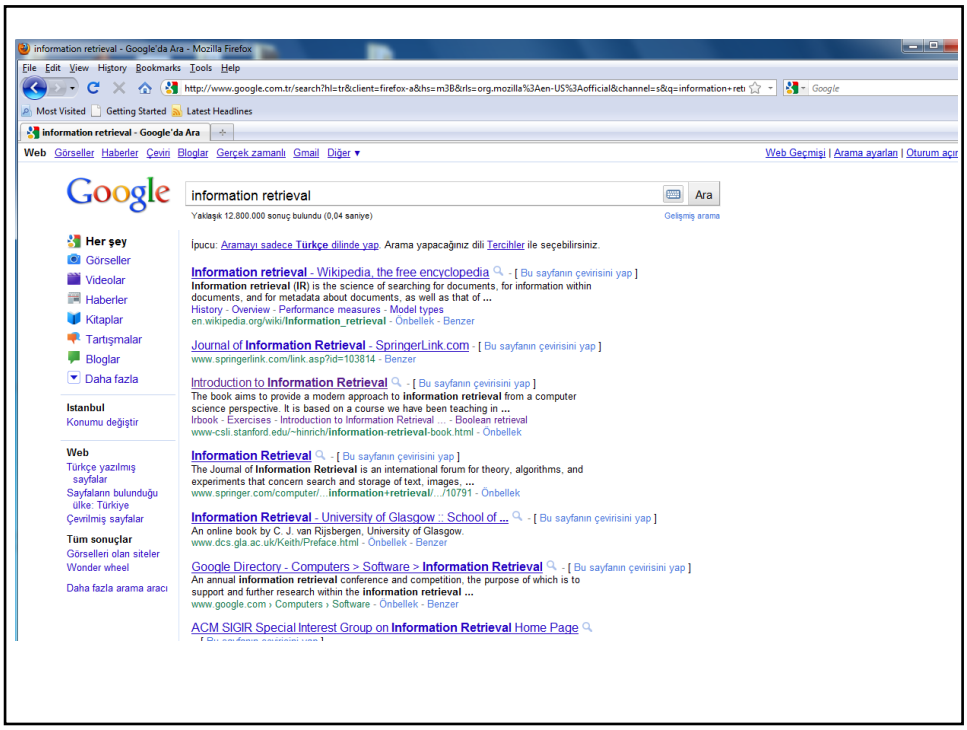


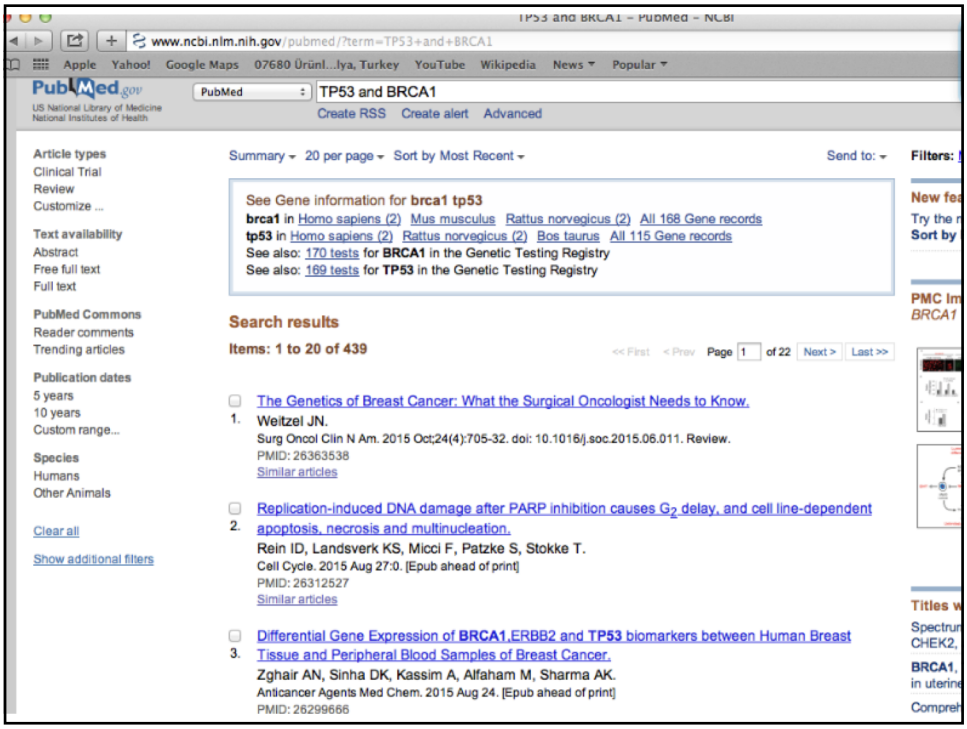


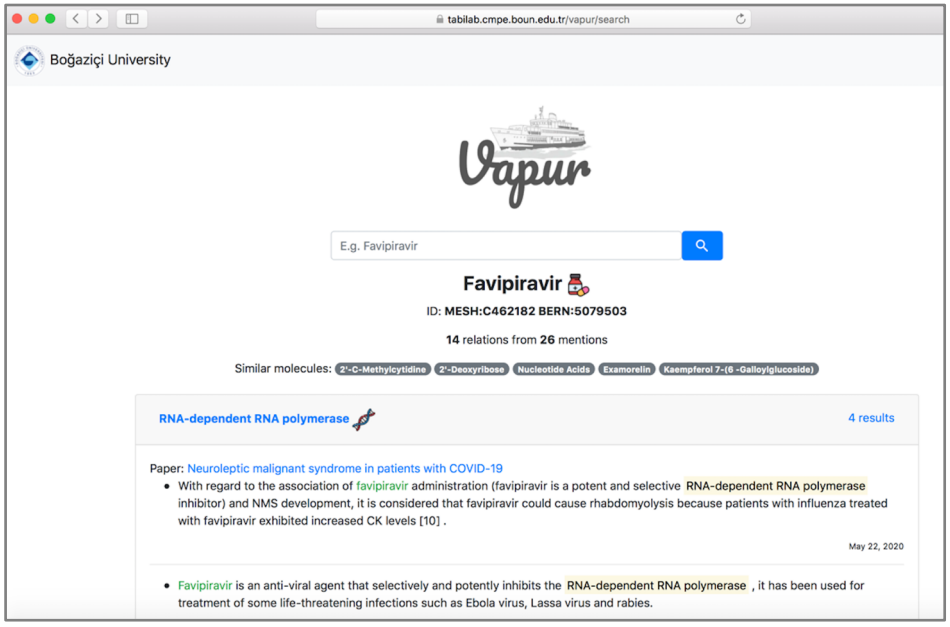
Laptop search
e.g. Spotlight

IR systems on the Web

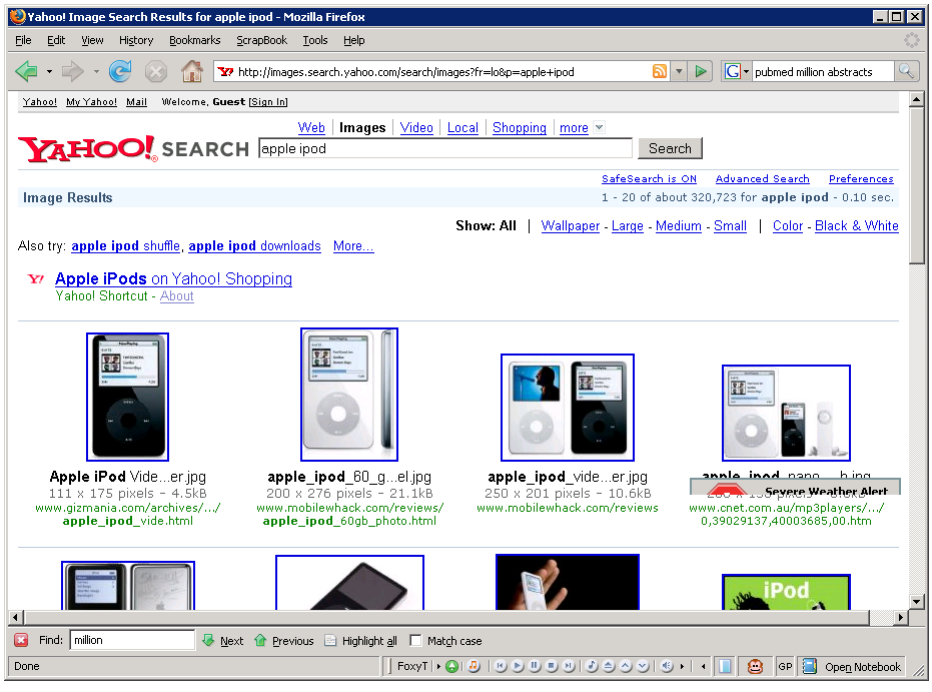
- Search for Web pages: http://www.google.com
- Domain specific search (e.g., legal, biomedical): PubMed
- Search for images: http://www.picsearch.com
- Search for image content: http://wang14.ist.psu.edu/
- Search for answers to questions: http://www.ask.com
- Music retrieval: http://www.rotorbrain.com/foote/musicr/

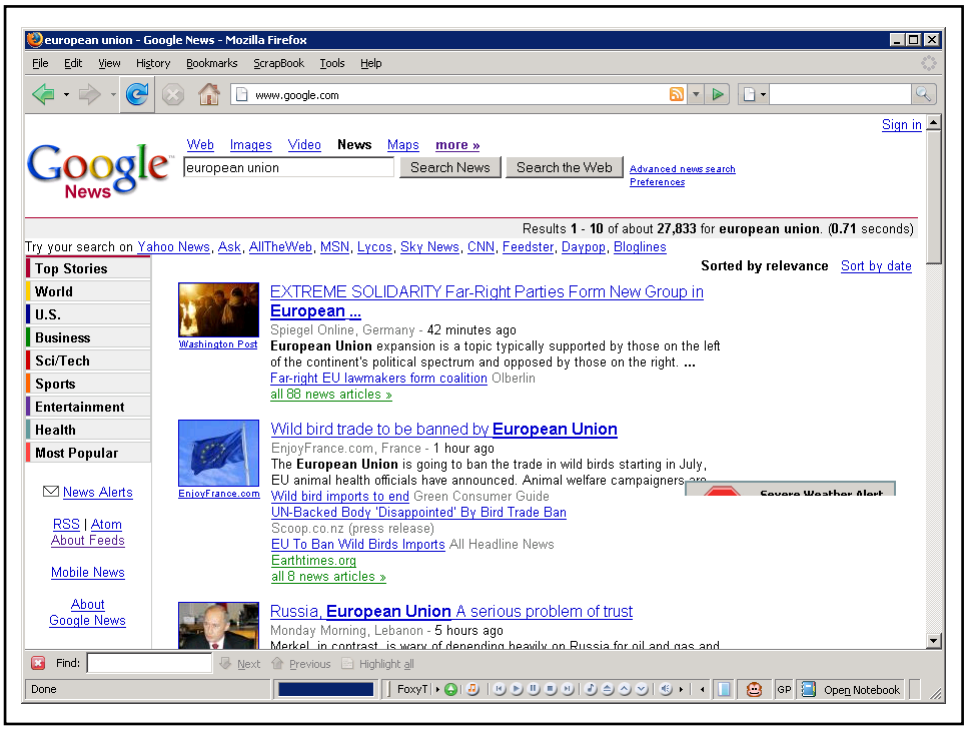

information retrieval - Google'da Ara - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.google.com.tr/search?hl=tr&client=firefox-a&hs=m3B&rls=org.mozilla%3Aen-US%3Aofficial&channel=s&q=information+ret
Most Visited Getting Started Latest Headlines
information retrieval - Google'da Ara
Web Görseller Haberler Çeviri Bloglar Gerçek zamanlı Gmail Diğer
Web Geçmişi | Arama ayarları | Oturum aç
Google
information retrieval
Ara
Yaklaşık 12.800.000 sonuç bulundu (0,04 saniye)
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
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

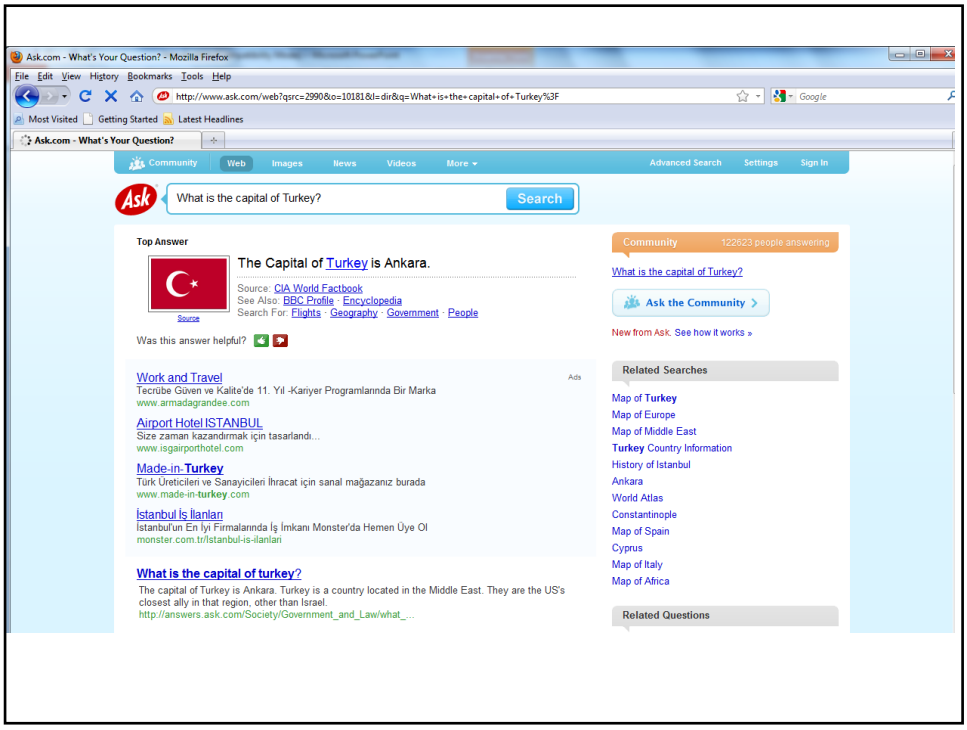

www.ncbi.nlm.nih.gov/pubmed/?term=TP53+and+BRCA1
Apple Yahoo! Google Maps 07680 Ürünl...lya, Turkey YouTube Wikipedia News Popular
PubMed.gov
US National Library of Medicine
National Institutes of Health
PubMed
TP53 and BRCA1
Create RSS Create alert Advanced
Article types
Clinical Trial
Review
Customize ...
Text availability
Abstract
Free full text
Full text
PubMed Commons
Reader comments
Trending articles
Publication dates
5 years
10 years
Custom range...
Species
Humans
Other Animals
Clear all
Show additional filters
Summary 20 per page Sort by Most Recent
Send to:
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 G2 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

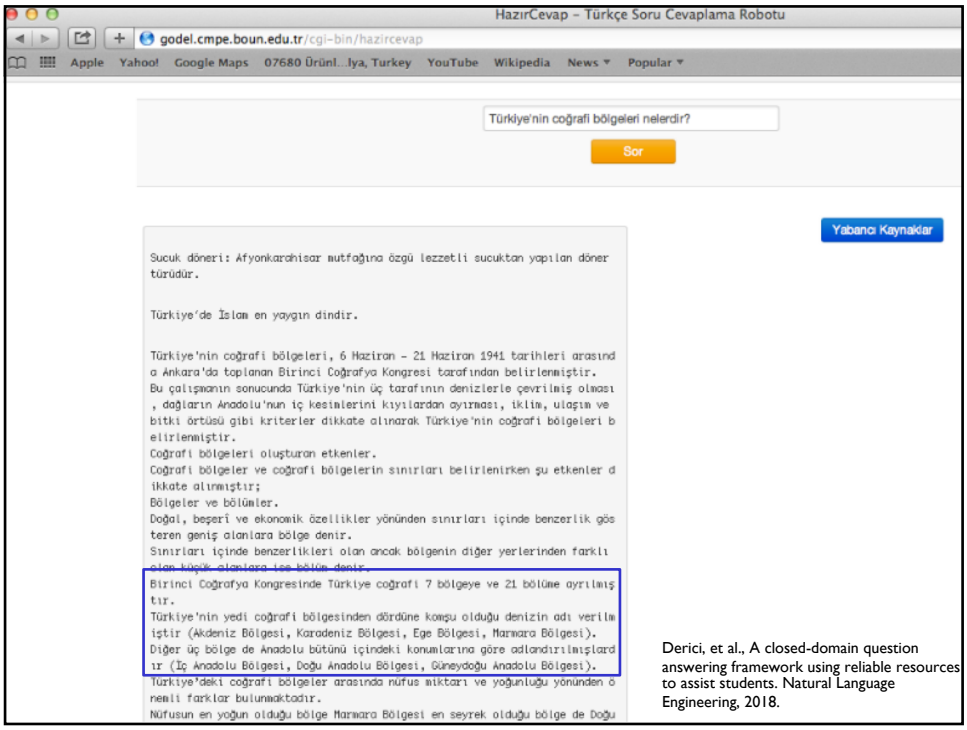


Köksal A, Dönmez H, Özçelik R, Ozkirimli E, Özgür A. **Vapur: A Search Engine to Find Related Protein - Compound Pairs in COVID-19 Literature.** *NLP COVID-19 Workshop, EMNLP 2020.*

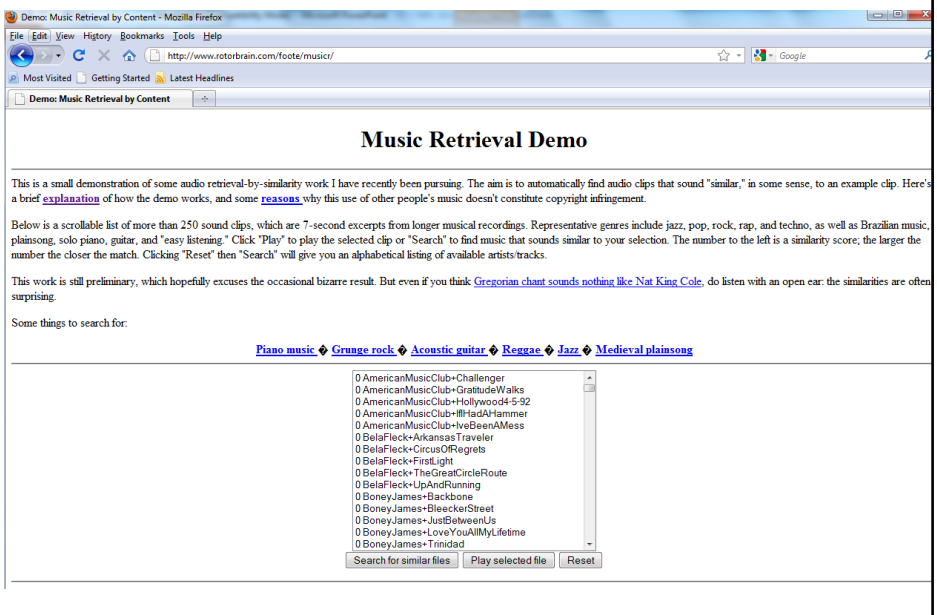



european union - Google News - Mozilla Firefox
File Edit View History Bookmarks ScrapBook Tools Help
www.google.com
Sign in
Google News
Web Images Video News Maps more »
european union
Search News
Search the Web
Advanced news search
Preferences
Results 1 - 10 of about 27,833 for european union. (0.71 seconds)
Try your search on Yahoo News, Ask, AllTheWeb, MSN, Lycos, Sky News, CNN, Feedster, Daypop, Bloglines
Top Stories
World
U.S.
Business
Sci/Tech
Sports
Entertainment
Health
Most Popular
News Alerts
RSS | Atom
About Feeds
Mobile News
About Google News
Sorted by relevance
Sort by date
EXTREME SOLIDARITY Far-Right Parties Form New Group in European ...
Washington Post
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 »
Wild bird trade to be banned by European Union
EnjoyFrance.com
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 »
Russia, European Union A serious problem of trust
Monday Morning, Lebanon - 5 hours ago
Find:
Next
Previous
Highlight all
Done
Open Notebook


Ask.com - What's Your Question? - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://www.ask.com/web?qsrc=2990&o=10181&l=dir&q=What+is+the+capital+of+Turkey%3F
Most Visited Getting Started Latest Headlines
Ask.com - What's Your Question?
Community Web Images News Videos More Advanced Search Settings Sign In
Ask
What is the capital of Turkey?
Search
Top Answer
The Capital of Turkey is Ankara.
Source: CIA World Factbook
See Also: BBC Profile · Encyclopedia
Search For: Flights · Geography · Government · People
Source
Was this answer helpful?
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?
Ask the Community
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



Derici, et al., A closed-domain question answering framework using reliable resources to assist students. Natural Language Engineering, 2018.



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 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
- 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
- Probabilistic models for information retrieval
- Text classification and clustering
- Latent semantic indexing
- Word embeddings for IR
- Web search and crawling
- Link analysis (e.g. hubs and authorities, Google PageRank)

References

- Some of the content adapted from Prof. Dragomir Radev and the IR book's web site.