

Web Search: Techniques, algorithms and Applications

Basic Techniques for Web Search

German Rigau <german.rigau@ehu.es>

[Based on slides by Eneko Agirre ...
and Christopher Manning and Prabhakar Raghavan]



Resources

- Information Retrieval book
- Possible coursework

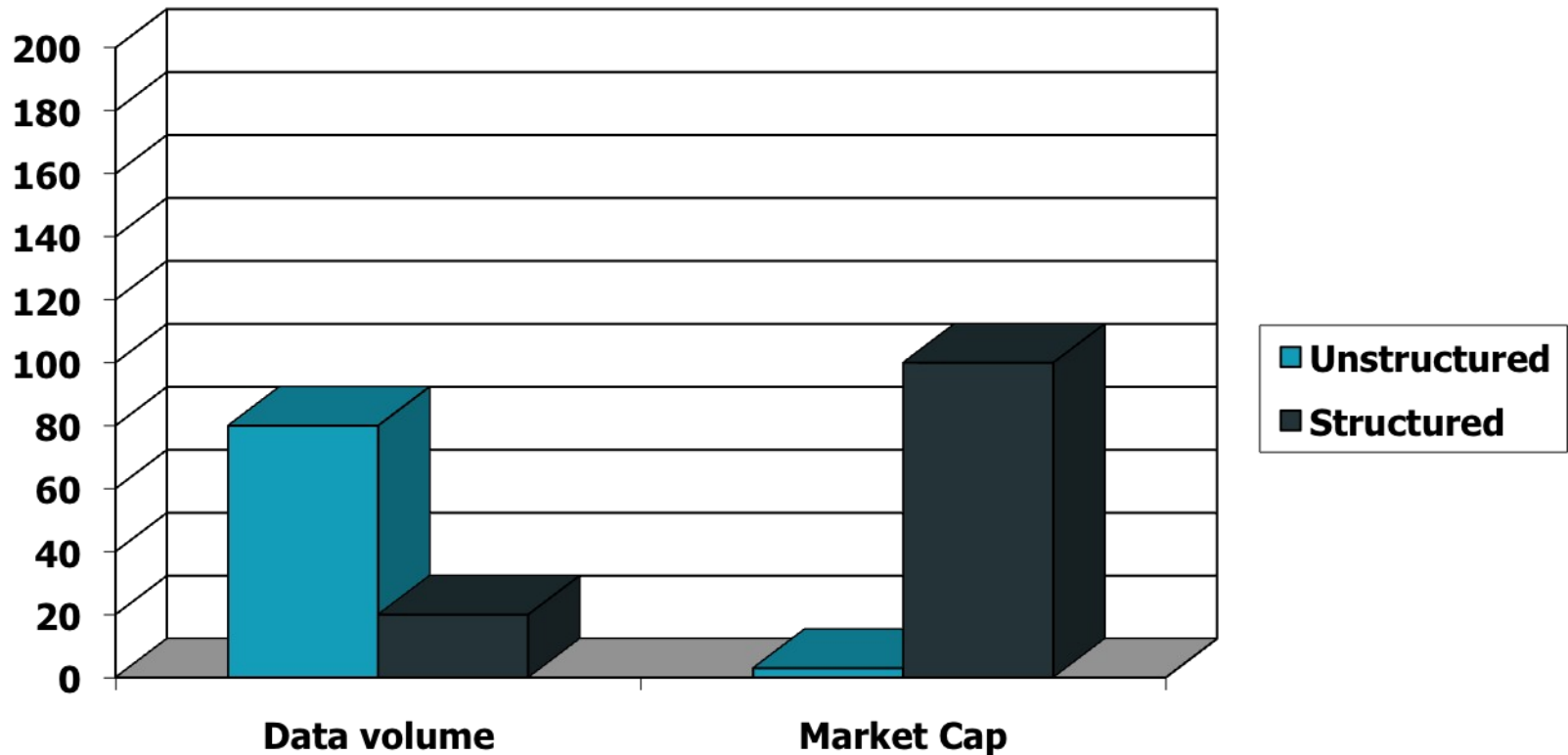
Basic Techniques for Web Search

- **Review of applications**
- Basic Techniques in detail
 - Boolean search
 - Vocabularies, dictionaries, index
 - Scoring, evaluation, complete system
 - Web search
- Semantic search

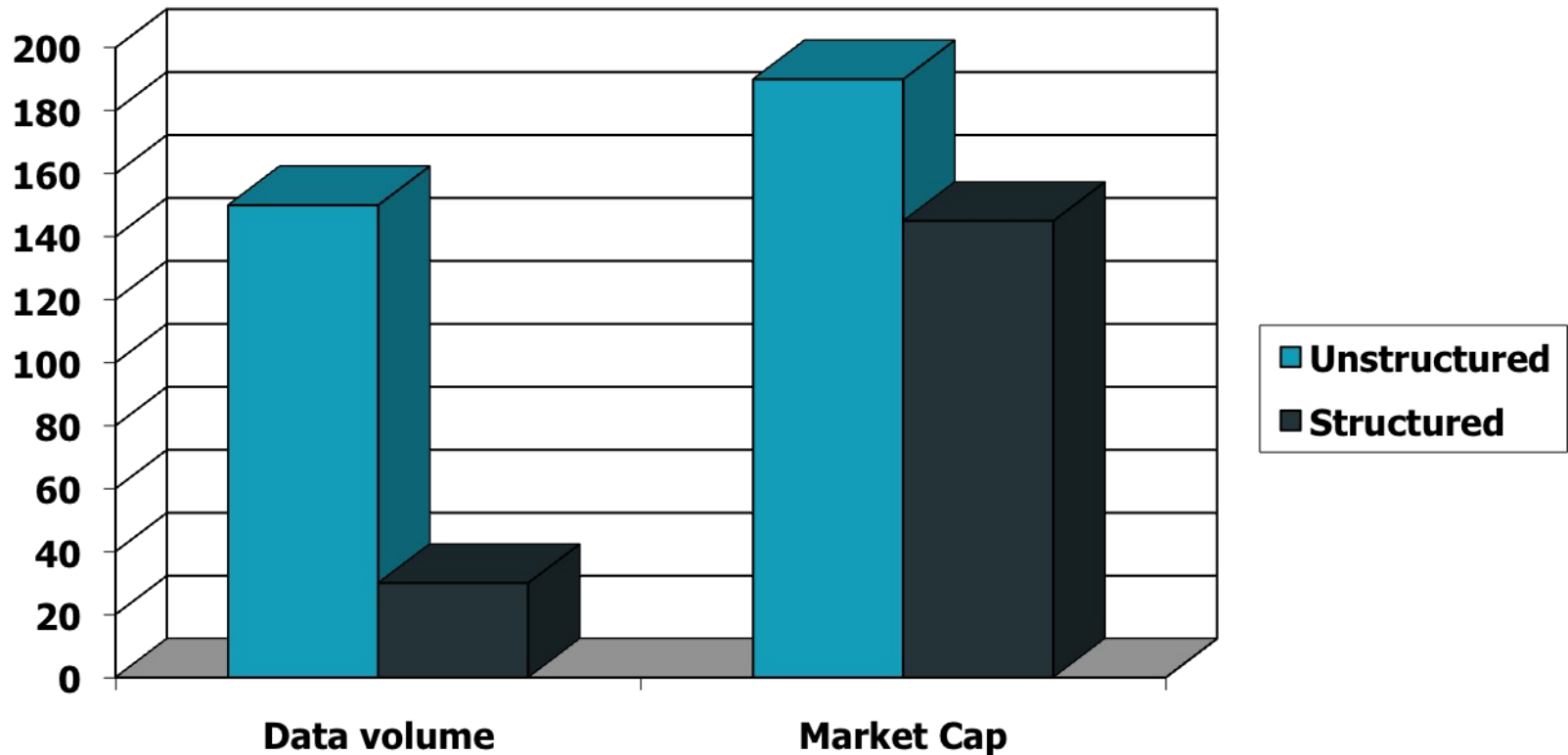
Information Retrieval

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

Unstructured (text) vs. structured (database) data in the 90s



Unstructured (text) vs. structured (database) data now



IR vs. databases:

Structured vs unstructured data

Structured data tends to refer to information in “tables”

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

Typically allows numerical range and exact match (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 dealing with *drug abuse*
- Classic model for searching text documents

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
 - XML search

... to say nothing of linguistic structure (NLP)

Semi-structured data



CC BY-NC

View item at

[The Wellcome Library](#)

Share

Cite on Wikipedia

Translate details

Select language ▼

Powered by **Microsoft**® Translator

A group of physicians stand by a cow that has died while 'ab

Description: A group of physicians stand by a cow that has died while 'aborting' a man; perhaps representing the 'aberration' of vaccination. Coloured etching.

Coverage: Paris (Rue de Coq) :

Type: Historical Images

Subject: COWS; divining-rod; elixir of life; ic; SATIRE; satire: french; Oxygen

Identifier:

<http://wellcomeimages.org/indexplus/miopac/V0011694.html>;

V0011694; ICV No 11959;

<http://catalogue.wellcomelibrary.org/record=b1174041>

Publisher: Michel

Source: V0011694

Data provider: The Wellcome Library

Provider: The European Library

Providing country: Europe



Review of applications

- Information Retrieval (IR)
 - Cross-lingual Information Retrieval (CLIR)
 - Question Answering (Q/A)
- Related tasks
 - Classification
 - Information Extraction (IE)
 - Summarization
 - Machine Translation

Need for IR

- With the advance of WWW -
<http://www.worldwidewebsize.com/>
- Corporate Intranets
- Personal PC's
- Various needs for information:
 - Search for (new) documents that fall in a given topic
 - Search for a specific (new) information
 - Search a (new) answer to a question
 - Search for information in a different language
 - ...
 - Search for images!
 - Search for music!
 - Search for a (candidate) friend ... similar hobbies, etc.

A definition of IR

Salton (1989): “Information-retrieval systems process files of records and requests for information, and identify and retrieve from the files certain records in response to the information requests. The retrieval of particular records depends on the similarity between the records and the queries, which in turn is measured by comparing the values of certain attributes to records and information requests.”

Examples of IR systems

- Conventional ([library catalog](#))
 - Search by keyword, title, author, etc.
- Text-based (Google, Yahoo, Bing ...).
 - Search by keywords. Limited search using queries in natural language.
 - Categorized search ([dmoz](#), [google directories](#))
 - Intranets ([ehu.es](#) - reglamento doctorado)
- Multimedia
 - Google [images](#) (it's mostly text)
 - By content (this is more semantic-web)
 - [Images](#)
 - Music (Shazam)
- Other:
 - Cross language information retrieval: [Elhuyar](#)
 - Question answering systems:
 - How many people live in New York? ... state?
 - What is the population of New York?

Review of applications

- Information Retrieval (IR)
 - Cross-lingual Information Retrieval (CLIR)
 - Question Answering (Q/A)
- Related tasks
 - Classification
 - Information Extraction (IE)
 - Summarization
 - Machine Translation

Document Classification

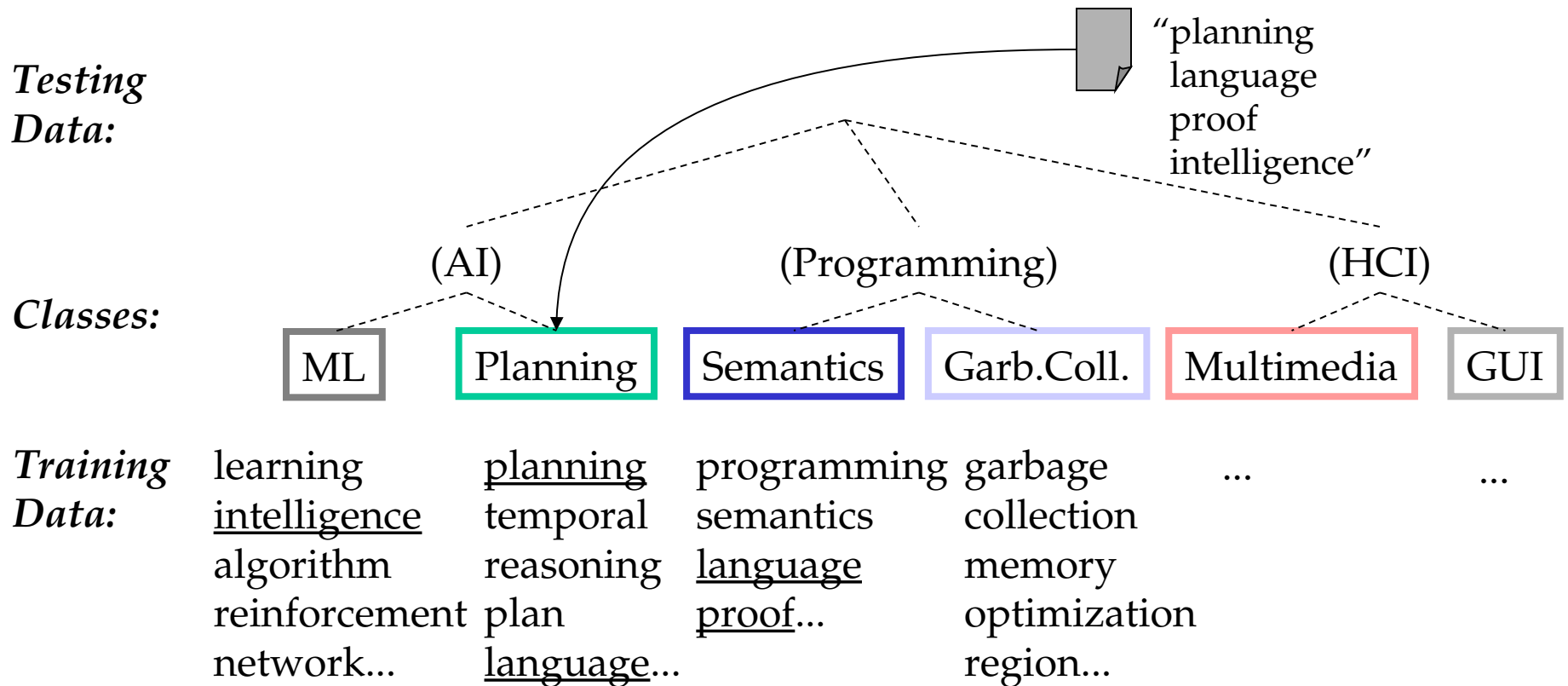
Assign labels to each document or web-page:

- Labels are most often **topics** such as Dmoz or Amazon
e.g., "finance," "sports," "news>world>asia>business"
- Labels may be **genres**
e.g., "editorials" "movie-reviews" "news"
- Labels may be **opinion**
e.g., "like", "hate", "neutral"
- Labels may be **domain**-specific binary
e.g., "interesting-to-me" : "not-interesting-to-me"
e.g., "spam" : "not-spam"
e.g., "is a toner cartridge ad" : "isn't"

Document Classification

- Given:
 - A description of an instance, $x \in X$, where X is a document.
 - Issue: how to represent text documents.
 - A fixed set of categories:
$$C = \{c_1, c_2, \dots, c_n\}$$
- Determine:
 - The category of x : $c(x) \in C$, where $c(x)$ is a *categorization function* whose domain is X and whose range is C .
 - We want to know how to build categorization functions (“classifiers”).

Document Classification



(Note: in real life there is often a hierarchy, not present in the above problem statement; and you get papers on ML approaches to Garb. Coll.)

Information Extraction (IE)

- Identify specific pieces of information (data) in a unstructured or semi-structured textual document.
- Transform unstructured information in a corpus of documents or web pages into a structured database.
- Applied to different types of text:
 - Newspaper articles
 - Web pages
 - Scientific articles
 - Newsgroup messages
 - Classified ads
 - Medical notes
 - Attributes of Named Entities (products, people, ...)

IE: Sample Job Posting

Subject: **US-TN**-SOFTWARE PROGRAMMER

Date: **17 Nov 1996** 17:37:29 GMT

Organization: Reference.Com Posting Service

Message-ID: <**56nigp\$mrs@bilbo.reference.com**>

SOFTWARE PROGRAMMER

Position available for Software Programmer experienced in generating software for PC-Based **Voice Mail** systems. Experienced in **C Programming**. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer **5 years** or more experience with **PC Based Voice Mail**, but will consider as little as **2 years**. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is **DOS**. May go to **OS-2** or **UNIX** in future.

Please reply to:

Kim Anderson

AdNET

(901) 458-2888 fax

kimander@memphisonline.com

IE: Extracted Job Template

computer_science_job
id: 56nigp\$mrs@bilbo.reference.com
title: SOFTWARE PROGRAMMER
salary:
company:
recruiter:
state: TN
city:
country: US
language: C
platform: PC \ DOS \ OS-2 \ UNIX
application:
area: Voice Mail
req_years_experience: 2
desired_years_experience: 5
req_degree:
desired_degree:
post_date: 17 Nov 1996

IE: Amazon Book Description

```
....
</td></tr>
</table>
<b class="sans">The Age of Spiritual Machines : When Computers Exceed Human Intelligence</b><br>
<font face=verdana,arial,helvetica size=-1>
by <a href="/exec/obidos/search-handle-url/index=books&field-author=
    Kurzweil%2C%20Ray/002-6235079-4593641">
Ray Kurzweil</a><br>
</font>
<br>
<a href="http://images.amazon.com/images/P/0140282025.01.LZZZZZZZ.jpg">
</a>
<font face=verdana,arial,helvetica size=-1>
<span class="small">
<span class="small">
<b>List Price:</b> <span class=listprice>$14.95</span><br>
<b>Our Price: <font color=#990000>$11.96</font></b><br>
<b>You Save:</b> <font color=#990000><b>$2.99 </b>
(20%)</font><br>
</span>
<p> <br>...
```

IE: Extracted Book Template

Title: The Age of Spiritual Machines :
When Computers Exceed Human Intelligence

Author: Ray Kurzweil

List-Price: \$14.95

Price: \$11.96

IE: real state

- <http://www.nuroa.es/>

Código Globaliza: 1470226

Características específicas

- > 103 m² útiles
- > 3 dormitorios
- > Para entrar a vivir
- > 2 baños
- > Puerta de Seguridad
- > Cocina totalmente equipada
- > Agua Caliente Central
- > Calefacción Central
- > Terraza de 12 metros
- > Trastero
- > Tendedero Cubierto
- > Gastos comunidad: 50.0€
- > 3 puertas por planta
- > Ascensor



Equipamiento

- > 1 plazas de garaje
- > Conserje/Portero/Seguridad
- > Circuito Cerrado de Seguridad
- > Edificio de 5 plantas
- > Antena Parabólica Colectiva

Observaciones

Piso en el Actur al lado Expo de 103.50m.3 habitacines + salon, cocina con galeria cerrada de 7 metros, salon de 22, terraza de 12 y dos baños completos uno en dormitorio principal.Armario emprotrado de 4 cuerpos, videoportero,urbanización privada,ventanas climalit,puertas roble,puerta blindada,exterior,calefacción central de gasoil,terrazo imitación a mármol.Garaje y trastero.Oportunidad, muy bien situado.Urge venta.350.000 euros.

IE: Scientific bibliography

- We can improve the library catalog with a citation index
- Match
 - Citation (Agirre & Rigau, 1996)
 - Reference: Agirre, E. and G. Rigau. 1996. Word sense disambiguation using conceptual density. Proceedings of COLING, 1996
 - Paper
- <http://scholar.google.com>
- <http://academic.research.microsoft.com>



what is the population of new york city

Search

About 41,500,000 results (0.17 seconds) [Advanced search](#)







- Everything
- Images
- Videos
- More

New York City population — 8,008,278 - [Feedback](#)
According to [infoplease.com](#), [nyc.gov](#), [demographia.com](#) and 1 other - [Hide sources](#)
[population of new york in ...](#) - [trueknowledge.com](#) - **8008278** is the 2000 US census population of New ...
[New York City Metropolitan ...](#) - [demographia.com](#) - CITY, **8008278**, 8164706, 8143197, (808552), 511018 ...
[Top 50 Cities in the U.S. ...](#) - [infoplease.com](#) - New York, N.Y., 8391881, 8143197, **8008278**, 7322564 ...
[Population - New York City ...](#) - [nyc.gov](#) - New York City, **8008278**, 8391881, 383603, 4.8 ...

► [New York City - Wikipedia, the free encyclopedia](#) 🔍
The **city's** 2009 estimated **population** approached 8.4 million, and with a land area of 305 square miles (790 km2), **New York City** is the most densely populated ...
[Borough - Demographics - Neighborhoods - Transportation](#)
[en.wikipedia.org/wiki/New_York_City](#) - [Cached](#) - [Similar](#)

[New York metropolitan area - Wikipedia, the free encyclopedia](#) 🔍
However, the **New York City** television designated market area (DMA) includes ...
[Components of the metropolitan ... - Geography - Transportation](#)
[en.wikipedia.org/wiki/New_York_metropolitan_area](#) - [Cached](#) - [Similar](#)

digital cameras

	Item Name <input type="button" value="v"/>	Image <input type="button" value="x"/>	Description <input type="button" value="x"/>	Weight <input type="button" value="v"/> <input type="button" value="x"/>	Dimensions <input type="button" value="v"/> <input type="button" value="x"/>	View <input type="button" value="x"/>
<input type="button" value="x"/>	Nikon Coolpix P50		The Coolpix P50 is Nikon's attempt at a middle-ground camera: it's neither a point-and-shoot-only ultra compact nor a fully-fledged photographers' camera. ...	5.6 oz	1.7 in.	Re...
<input type="button" value="x"/>	Nikon Coolpix 950		Since the story broke back in mid--February, the Nikon Coolpix 950 (referred to as the 950 from now onwards) has generated a fever which has gripped the ...	0.4 Kg	1.4 in.	Op...
<input type="button" value="x"/>	Sony Cyber-shot DSC-W80		While the Sony Cyber-shot DSC-W80 has quite a few things going for it (such as image stabilization and robust performance), the negatives outweigh the ...	4.4 oz	3.6 x 2.3 x 0.9 in. www.dcresource.com Choose from 4 values »	Re...
<input type="button" value="x"/>	Ricoh Caplio GX100		Announced back in March the Ricoh Caplio GX100 is officially the successor to the GX8, but it could equally well be described as a zoom version of the ...	220 g..	25 mm	Ex...
<input type="button" value="x"/>	Canon Powershot SD990 IS		When you're ready to experience a higher level of pace-setting technology and image brilliance, the PowerShot SD990 IS Digital ELPH is ready for you. ...	5.6 oz	1.10 in.	Op...
<input type="button" value="x"/>	Vivitar		Vivitar - effortless experience. Search. Copyright Sakar Inc. 2009 I disclaimer	0.22 lb	2.2 in.	

Named Entity Disambiguation (aka Entity linking)

Given Knowledge Base *string* **Paul Newman** *entity* **E0181364**

Given target string and surrounding text:

I watched “Slapshot”, the 1977 hockey classic starring Paul Newman for the first time.

Return entity in KB (**E0181364**) or NIL

KB subset/superset of Wikipedia

Paul Newman

E0181364



The image shows a screenshot of a web browser displaying the Wikipedia disambiguation page for "Paul Newman". The browser's address bar shows the URL "http://en.wikipedia.org/wiki/Paul_Newman". The page title is "Paul Newman (disambiguation)". The main content area lists several disambiguation options: "Paul Newman (politician)", "Paul Newman (cricketer)", "Paul Newman (linguist)", and "Paul Newman (band)". The page also includes a "See also" section and a "Contents" table of contents. The browser's address bar shows the URL "http://en.wikipedia.org/wiki/Paul_Newman". The page title is "Paul Newman (disambiguation)". The main content area lists several disambiguation options: "Paul Newman (politician)", "Paul Newman (cricketer)", "Paul Newman (linguist)", and "Paul Newman (band)". The page also includes a "See also" section and a "Contents" table of contents.

Wikipedia disambiguation page for Paul Newman. The page lists several disambiguation options:

- Paul Newman (politician), Arizona politician
- Paul Newman (cricketer), English cricketer
- Paul Newman (linguist), American linguist
- Paul Newman (band), an Austin, Texas band

The page also includes a "See also" section and a "Contents" table of contents.

Contents [hide]
1 Early life
1.1 Military service
2 Career
2.1 Early work
2.2 Major films
2.3 Last works
2.4 Retirement from acting
3 Philanthropy

Slot filling (Infobox generation)

Distant supervision (Mintz et al. 09):

- Use facts in Knowledge Base
=> **gold-standard entity–slot–filler**
- Search for spans containing entity–filler pair in document base
=> **positive examples to train**
- Train classifier per slot
- Search for mentions of target entity in document collection
- Run each of the classifiers

Manual work kept to a minimum.

Paul Newman



in 2007

Born	Paul Leonard Newman January 26, 1925 Shaker Heights, Ohio, U.S.
Died	September 26, 2008 (aged 83) Westport, Connecticut, U.S.
Occupation	Actor, director, humanitarian, entrepreneur
Years active	1952–2007
Spouse(s)	Jackie Witte (1949–1958) (divorced) Joanne Woodward (1958–2008) (his death)

Google Knowledge Graph

Paul Newman - Google Search

www.google.com/search?q=lll&ie=UTF-8&sa=Search&channel=fe&client=browser-ubuntu&hl=en#hl=en&sugexp=les%3B&g

+You Search Images Maps Play YouTube News Gmail Documents Calendar More -

Google

Sign in

Web Images Maps Shopping News More Search tools

About 30,200,000 results (0.30 seconds)

Paul Newman - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Paul_Newman

Paul Leonard Newman (January 26, 1925 – September 26, 2008) was an American actor, film director, entrepreneur, humanitarian, professional racing driver, ...

Joanne Woodward - Scott Newman (actor) - Newman's Own - Cool Hand Luke

Paul Newman - IMDb

www.imdb.com/name/nm0000056/


Paul Newman, Actor: Butch Cassidy and the Sundance Kid. Paul Leonard Newman was born in January of 1925, the second son of Arthur and Theresa (nee' ...

Newman's Own

www.newmansown.com/

The Official Website of Newman's Own. ... Sign up for Newsletter · Life & Legacy of Paul Newman · Newman's Own Foundation · SeriousFun Children's Network ...

Images for paul newman - Report images




Paul Newman, a Magnetic Titan of Hollywood, Is Dead at 83 ...

www.nytimes.com/2008/09/28/movies/28newman.html?...all

27 Sep 2008 – Mr. Newman, one of the last of the great 20th-century stars, acted in more than 65 movies over half a century.

Paul Newman



imdb.com

Paul Leonard Newman was an American actor, film director, entrepreneur, humanitarian, professional racing driver, auto racing team owner, and auto racing enthusiast. Wikipedia

Born: January 26, 1925, Shaker Heights


Died: September 26, 2008, Westport

Height: 1.77 m

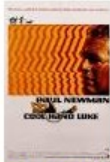
Spouse: Joanne Woodward (m. 1958–2008), Jackie Witte (m. 1949–1958)

Children: Scott Newman, Nell Newman, Melissa Newman, Claire Olivia Newman, More

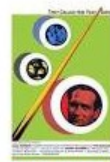
Movies and TV shows




Butch Cassidy and the S...
1969




Cool Hand Luke
1967



The Hustler
1961




The Sting
1973



The Color of Money
1986

People also search for



Google Knowledge Graph



Web Images Maps Shopping More Search tools

About 113,000,000 results (0.24 seconds)

[Colorado Vacation: Colorado Sightseeing - Durango Area Tourism ...](#)

www.durango.org/

The **Durango** Area Tourism Office offers a variety of information about Colorado tourist attractions, **Durango** hotels, and everything you need to plan your next ...

[Durango - Wikipedia, the free encyclopedia](#)

en.wikipedia.org/wiki/Durango

Durango officially Free and Sovereign State of **Durango** (Spanish: Estado Libre y Soberano de **Durango**), is one of the 31 states which, with the Federal District, ...

[Durango, Durango](#) - [Durango, Biscay](#) - [Gómez Palacio](#) - [Ciudad Lerdo](#)

[Durango, Colorado Vacation Guide: Official Tourism & Travel guide ...](#)

www.durango.com/

See results about



Durango

Colorado

The City of Durango is a Home Rule Municipality that is the county seat and the most populous city of La ...



Durango

Durango, officially Free and Sovereign State of Durango, is one of the 31 states which, with the ...



durango

durango – City in Colorado

durango – Mexican state

durango herald

durango boots

durango mountain resort








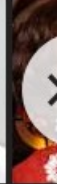
Press Enter to search.

Google Knowledge Graph

Google

Web Images Maps Shopping More ▾ Search tools

Avatar cast

							
Zoe Saldana Neytiri	Sam Worthington Jake Sully	Michelle Rodriguez Trudy Chacon	Sigourney Weaver Grace	Stephen Lang Colonel Miles Quaritch	Giovanni Ribisi Parker Selfridge	Laz Alonso Tsu'Tey	C. O'Connell Mo'at

[Avatar \(2009\) - Full Cast & Crew - IMDb](http://www.imdb.com/title/tt0499549/fullcredits)

www.imdb.com/title/tt0499549/fullcredits ▾

Sam Worthington ... Jake Sully · Zoe Saldana ... Neytiri (as Zoë Saldana). Sigourney Weaver ... Grace · Stephen Lang ... Colonel Miles Quaritch · Michelle ...

[Images for what is the cast of](#) - Report images



[Avatar \(2009 film\) - Wikipedia, the free encyclopedia](http://en.wikipedia.org/wiki/Avatar_(2009_film))

[en.wikipedia.org/wiki/Avatar_\(2009_film\)](http://en.wikipedia.org/wiki/Avatar_(2009_film)) ▾

Jump to **Cast** - Further information: Fictional universe of **Avatar** ... Cameron cast the Australian actor after a worldwide search for promising young **actors**, ...

[Sam Worthington](#) - [Stephen Lang](#) - [Unobtainium](#) - [Fictional universe](#)

Avatar



2009 Film

Avatar is a 2009 American epic science fiction action film written and directed by James Cameron, and starring Sam Worthington, Zoe Saldana, Stephen Lang, Michelle Rodriguez, Joel David Moore, Giovanni Ribisi, and Sigourney Weaver.

[Wikipedia](#)

Release date: December 18, 2009 (USA)

Director: [James Cameron](#)

Sequel: [Avatar 2](#)

Budget: 237 million USD (2009)

People also search for

Summarization

lando2001_IXA.doc - Microsoft Word

Edit View Insert Format Tools Table Window Help

Spelling and Grammar... F7

Language

Word Count...

AutoSummarize...

AutoCorrect...

Track Changes

Merge Documents...

Protect Document...

Online Collaboration

Mail Merge...

Envelopes and Labels...

Letter Wizard...

Macro

Templates and Add-Ins...

Customize...

Options...

REC title

Times New Roman

18

B

I

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

☐

AutoSummarize



25%

Close

text, all its possible
of speech are shown
equivalent in the other

an a pure consult to the
not obvious which is the
orm. For example, in a
ok up the entry for the
the meaning of the
but as the entry *Caber*
cuníramos might be
e user could give up
equivalent, while our
offers the candidate
lemmas and their corresponding equivalents (see Figure 5).

4.3. A lemmatization based web-crawler

GAIN is a tool for Internet/Intranet search engine composed by a robot, an indexer and a client for queries. It contains two NLP modules that make it a powerful tool: a language identification module and the lemmatizer.

The Swish-E tool², from the University of Berkeley, has been adapted, as it provides the required features: modularity, completeness, freeware, multiplatform and multiformat. It has been installed on a Web Linux-Apache server.

The most important improvement we have made has been a lemma-based indexing. The lemmatizer has been integrated and, as a result, we have obtained both an advanced search engine and a more compact index. In

We have presented a long-term open proposal to develop language technology for lesser-used languages. We have defined four phases that involve parallel and coordinated creation of language foundations, tools and applications. The development of a sound language industry should be the result of a synchronized effort, involving research groups, institutions and industry. This way, language foundations and research will be useful to create any tool or application and, conversely, tools and applications will be very helpful in research and improving language foundations

Now that we have started working on the fourth phase, every foundation, tool and application developed in the previous phases is of great importance to face new problems. We have presented several commercial applications, including the spelling checker and the lemmatizer, which are very active tools in the standardization process of Basque.

Acknowledgements

This work has been supported by the Department of Economy of the Government of Gipuzkoa, The University of the Basque Country, the Department of Education of the Basque Government and the Commission of Science and Technology of the Spanish Government.

References

Aduiz I., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Goienola K., Maritxalar A., Sarasola K., Urkia M.] "A Word-grammar based

Machine Translation

Búsqueda en Google: J.K. Rowling - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites History Print Link

Address <http://www.google.es/search?hl=es&ie=ISO-8859-1&q=J.K.+Rowling+&btnG=B%FAsqveda&meta=> Go Links Customize Links

Google

La Web [Imágenes](#) [Grupos](#) [Directorio](#) [News ¡Nuevo!](#)

J.K. Rowling Búsqueda

[Búsqueda Avanzada](#)
[Preferencias](#)

Búsqueda: ☒ la Web ☐ páginas en español ☐ páginas de España

La Web Resultados 1 - 10 de aproximadamente 594.000 de J.K. Rowling . (0.18 segundos)

[JK.Rowling Official Site - Harry Potter and more](#) - [[Traduzca esta página](#)]
Jump aboard the Hogwart Express to come on a fantastic adventure at the official JK Rowling Website. See what's on my desk. ... All rights reserved JK Rowling.
[www.jkrowling.com/](#) - 4k - 21 Jun 2004 - [En caché](#) - [Páginas similares](#)

[Harry Potter: Meet JK Rowling](#) - [[Traduzca esta página](#)]
... JK Rowling has won the Hugo Award, the Bram Stoker Award, the Whitbread Award for Best Children's Book, a special commendation for the Anne Spencer Lindbergh ...
[www.scholastic.com/harrypotter/author/](#) - 23k - [En caché](#) - [Páginas similares](#)

[Harry Potter](#) - [[Traduzca esta página](#)]
The Scholastic Store. Select. ...
[www.scholastic.com/harrypotter/home.asp](#) - 16k - 21 Jun 2004 - [En caché](#) - [Páginas similares](#)
[[Más resultados de www.scholastic.com](#)]

[JK Rowling](#) - [[Traduzca esta página](#)]
Like that of her own character, Harry Potter, JK Rowling's life has the luster of a fairy tale. ... Courtesy of Book Magazine. An Interview with JK Rowling. ...
[www.kidsreads.com/harrypotter/jkrowling.html](#) - 26k - [En caché](#) - [Páginas similares](#)

Applications and tools: what is the relation?

- Information Retrieval (IR)
 - Conventional (library catalog)
 - Text-based
 - Multimedia by content
 - Cross-lingual Information Retrieval (CLIR)
 - Question Answering (Q/A)
- Related tasks
 - Information Extraction (IE)
 - Classification
 - Summarization
 - Machine Translation
- Tools
 - Tokenization
 - Sentence Splitting
 - Language Identifiers
 - Lemmatization, POS tagging
 - Named Entity Recognizers and Categorizers (NERC)
 - Parsing
 - Named Entity Dis. (NED)
 - Word Sense Dis. (WSD)
 - Semantic Role Labelling (SRL)

Applications and resources: what is the relation?

- Information Retrieval (IR)
 - Conventional (library catalog)
 - Text-based
 - Multimedia by content
 - Cross-lingual Information Retrieval (CLIR)
 - Question Answering (Q/A)
- Related tasks
 - Information Extraction (IE)
 - Classification
 - Summarization
 - Machine Translation
- Resources
 - Wordnets, EuroWordnets, MCR
 - Topic signatures
 - FrameNet
 - SUMO
 - CYC
 - ...
 - Domain ontologies: UMLS
 - Wikipedia / DBpedia /
Freebase /Linked-open-data

Web Search: Techniques, algorithms and Applications

Basic Techniques for Web Search

German Rigau <german.rigau@ehu.es>

[Based on slides by Eneko Agirre ...
and Christopher Manning and Prabhakar Raghavan]

